Final Report

Grant Title: Two-Dimensional Malvar Wavelets and Their Applications in Jamming Resistance Communication and Navigation

Grant Number: AFOSR # F49620-97-1-0253

Principal Investigator: Xiang-Gen Xia

Institution: University of Delaware

Reporting Period: 1 June 1997 – 31 December 1997

Award Period: 1 June 1997 – 31 December 1997

Submitted By:

Xiang-Gen Xia
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716
Phone/Fax: (302)831-8038/4316
Email: xxia@ee.udel.edu

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-00-

0329

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	7-1-1999	Final Report, 6-1-1997 -- 12-31-1997

4. TITLE AND SUBTITLE
Two-Dimensional Malvar Wavelets and Their Applications in Jamming Resistance Communication and Navigation

5. FUNDING NUMBERS
G F49620-97-1-0253

6. AUTHOR(S)
Xiang-Gen Xia

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716

8. PERFORMING ORGANIZATION REPORT NUMBER
UODECE. SF298REPORT 1-7-99

9. SPONSORING / MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Air Force Office of Scientific Research

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

a. DISTRIBUTION / AVAILABILITY STATEMENT

12. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report describes the main research achievements during the time period cited above on the research project in the area of signal processing and telecommunications. The main achevements include the construction of Malvar wavelets on arbitrary shapes, a new system identification method using chirp signals and joint time-frequency analysis method, a new prefiltering for discrete multiwavelet transforms, and some new signal processing methods for telecommunications and radar applications of jamming resistance.

14. SUBJECT TERMS
Malvar wavelets, system identification, anti-jamming, radar applications, DFT, watermarking, multiwavelets

15. NUMBER OF PAGES
6

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

A. **Objective**: The goal of this research is to develop new and better digital communication systems using wavelets and multirate filterbanks, which include new source and channel codings for antijamming applications.

B. **Main Research Accomplishments**: We have made several research accomplishments as follows.

(i). *Malvar Wavelets on Arbitrary Shapes.* We systematically constructed two dimensional Malvar wavelets defined on L-shaped regions, which can be used to construct two dimensional Malvar wavelets on arbitrary shapes. The application of such wavelets include the discrete cosine transform (DCT) image coding for arbitrary shaped objects such as the emerged video coding standard MPEG4. The two dimensional Malvar wavelets may be used the eliminate the block effects produced by the DCT compression at high compression ratios.

(ii). *Ambiguity Resistant Precoders for ISI Mitigation Without Channel Information.* We developed new precoding schemes using multirate filterbanks, which are called ambiguity resistant precoders (ARP). With ARP, the ISI channel information is not necessary for the transmitter or the receiver to recover the information. We characterized all ARP, which are some special families of matrix polynomials. Any ARP can be used to resist ISI, to also resist additive random errors optimal ARP were studied and characterized.

(iii). *A New System Identification Method.* New channel identification using chirp signals and joint time-frequency analysis and synthesis was proposed, where the performance is superior to the conventional method at low SNR.

(iv). *Multiwavelet Transforms.* A new prefiltering for discrete multiwavelet transforms, which has better energy compaction than other prefiltering, was obtained.

(v). A family of new pulse shaping filters that are ISI free with the matched filtering and also ISI free wothout the matched filtering, was obtained, which has been recently extended by numerous researchers.

(vi). A quantatitive SNR analysis for joint time-frequency analysis by introducing 3dB SNR definition in the joint time-frequency plane was obtained.

(vii). A frequency estimation method from the undersampled data with multiple frequencies was proposed.

(viii). An optimal multiple pulse repetition frequency (PRF) design method was proposed.

(ix). A new wavelet based watermarking was proposed.

C. **Significance**: The results obtained through this project have advanced digital signal processing and its applications, in particular wavelets and filterbanks, in telecommunications, multimedia systems, and radar applications with some jamming resistance properties.

D. **Publications, Abstracts, Technical Reports, and Patent Disclosures or Applications**

Published and Accepted (Refereed) Journal Publications

1. G. Zhou and X.-G. Xia, Multiple frequency detection in undersampled complex-valued waveforms with close multiple frequencies, *Electronics Letters*, vol.33, no.15, pp.1294-1295, July 1997.
2. X.-G. Xia, System identification using chirp signals and time-variant filters in the joint time-frequency domain, *IEEE Trans. on Signal Processing*, vol. 45, pp.2072-2084, August, 1997.
3. X.-G. Xia, A family of pulse shaping filters with ISI-free matched and unmatched filter properties, *IEEE Trans. on Communications*, vol. 45, Oct. 1997.
4. X.-G. Xia and M.Z. Nashed, A method with error estimates for band-limited signal extrapolation from inaccurate data, *Inverse Problems*, Dec., 1997.
5. X.-G. Xia, A quantitative analysis of SNR in the short-time Fourier transform domain for multicomponent signals, *IEEE Trans. on Signal Processing*, vol.46, Jan. 1998.
6. X.-G. Xia, Smooth local sinusoidal bases on two dimensional L-shaped regions, *The Journal of Fourier Analysis and Applications*, vol.4, Issue 1, pp.53-66, 1998.
7. X.-G. Xia, A new prefilter design for discrete multiwavelet transforms, *IEEE Trans. on Signal Processing*, vol.46, June 1998.
8. X.-G. Xia, Orthonormal matrix valued wavelets and matrix Karhunen-Loeve expansion, *Contemporary Mathematics*, vol.216, 1998.
9. X.-G. Xia, C. G. Boncelet, and G. R. Arce, A wavelet transform based watermark for digital images, *Optical Express*, Dec. 1998.

10. X.-G. Xia, Doppler ambiguity resolution using optimal multiple pulse repetition frequencies, *IEEE Trans. on Aerospace and Electronic Systems*, vol. 35, Jan. 1999.

11. G. Zhou and X.-G. Xia, Ambiguity resistant polynomial matrices, *Linear Algebra and its Applications*, vol. 286, pp.19-35, 1999.

12. X.-Q. Gao, Z.-Y. He, and X.-G. Xia, Efficient implementation of arbitrary-length cosine-modulated filter banks, *IEEE Trans. on Signal Processing*, vol.47, April 1999.

13. X.-G. Xia and V. C. Chen, A quantitative SNR analysis for pseudo Wigner-Ville distributions, *IEEE Trans. on Signal Processing*, Oct. 1999, to appear.

14. X.-G. Xia and S. Qian, Convergence of an iterative time-variant filtering based on discrete Gabor transform, *IEEE Trans. on Signal Processing*, Oct. 1999, to appear.

15. X.-G. Xia, On estimation of multiple frequencies in undersampled complex valued waveforms, *IEEE Trans. Signal Processing*, to appear.

Conference Proceeding Publications

1. X.-G. Xia and S. Qian, An iterative algorithm for time-variant filtering in the discrete Gabor transform domain, Proceedings of IEEE ICASSP'97, Munich, Germany, April, 1997.

2. X.-G. Xia, Intersymbol interference cancellation using nonmaximally decimated multirate filterbanks, Proceedings of the New Jersey Institute of Tech. Symp'97 on Wavelet, Subband and Block Transforms in Communications, Newark, New Jersey, March 21, 1997.

3. X.-G. Xia, System identification in low SNR environment using chirp signals and time-variant filters, Conference on Information Sciences and Systems, John Hopkins University, March 19-21, 1997.

4. X.-G. Xia, A quantitative study of SNR for short-time Fourier transform, Conference on Information Sciences and Systems, John Hopkins University, March 19-21, 1997.

5. X.-G. Xia, A new precoding for ISI cancellation using multirate filterbanks, Proceedings of IEEE ISCAS'97, Hong Kong, June 9-12, 1997.

6. X.-G. Xia, Channel estimation in low SNR environments using chirp signals and joint time-frequency filters, The First IEEE Signal Processing Society Workshop on Multimedia Signal Processing, Princeton, New Jersey, June 23-25, 1997.

7. X.-G. Xia, C. G. Boncelet, and G. R. Arce, A multiresolution watermark for digital images, 1997 IEEE International Conf. on Image Processing, Santa Barbara, CA, Oct.26-29, 1997.

8. H. Liu and X.-G. Xia, Precoding for undersampled antenna array receiver systems, Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, Nov. 1997.

9. X.-G. Xia, A New Prefilter Design for Discrete Multiwavelet Transforms, Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, Nov. 1997.

10. X.-G. Xia and G. Zhou, Multiple Frequency Detection in Undersampled Waveforms, Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, Nov. 1997.

11. X.-G. Xia, Ambiguity Resistant Precoders in ISI/Multipath Cancellation: Distance and Optimality, Proceedings of the 28th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, Nov. 1997.

12. X.-G. Xia, Channel identification with Doppler and time shifts using mixed training signals, Proceedings of IEEE ICASSP'98, Seattle, May 12-15, 1998.

13. X. Gao, Z. He, and X.-G. Xia, A new implementation of arbitrary length cosine modulated filterbanks, Proceedings of IEEE ICASSP'98, Seattle, May 12-15, 1998.

14. X.-G. Xia, Doppler ambiguity resolution using optimal multiple pulse repetition frequencies, Conference on Information Sciences and Systems, Princeton University, March 18-20, 1998.

15. X.-G. Xia, Dynamic range determination of the detectable parameters for polynomial phase signals using multiple lag diversities in high-order ambiguity functions, Conference on Information Sciences and Systems, Princeton University, March 18-20, 1998.

16. G. Wang, X.-G. Xia and V. Chen, Signal-to-noise analysis in the joint time-frequency analysis domain for ISAR imaging, Proceedings of SPIE'98, San Diego, July 20-25, 1998.

E. Collaborators

Dr. Hui Liu at University of Washington, Seattle, WA.

Dr. Victor C. Chen at Naval Research Laboratory.

Mr. Shie Qian at National Instruments.

F. Post Doctors and Students Supported and Partially Supported by the Grant

Post Doctor:

Dr. Guangcai Zhou

Student:

Mr. Kai Bao

G. Appendix: Journal Publications

1 NARAHASHI, S., and NOJIMA, T.: 'New phasing scheme of N-multiple carriers for reducing peak-to-average power ratio', *Electron. Lett.*, 1994, **30**, pp. 1382–1383
2 LI, X., and RITCEY, J.A.: 'M-sequences for OFDM peak-to-average power ratio reduction and error correction', *Electron. Lett.*, 1997, **33**, pp. 554–555
3 GOLOMB, S.W.: 'Shift register sequences' (Aegean Park Press, California, 1982), revised edn.
4 DAVIS, J.A., and JEDWAB, J.: 'Peak-to-mean power control and error correction for OFDM transmission using Golay sequences and Reed-Muller codes', *Electron. Lett.*, 1997, **33**, pp. 267–268

Multiple frequency detection in undersampled complex-valued waveforms with close multiple frequencies

Guangcai Zhou and Xiang-Gen Xia

Indexing terms: Signal processing, Frequency measurement

The determination of multiple frequencies in undersampled waveforms is studied, where the multiple frequencies are close to each other. Given multiple undersampling rates, the maximal range of the detectable multiple frequencies is the least common multiple of these multiple rates under the assumption that these rates are larger than twice the maximal distance between the multiple frequencies.

Introduction: A traditional method to detect the frequencies of a multiple frequency waveform $x(t)$ is to sample $x(t)$ at the Nyquist sampling rate and then to implement the discrete Fourier transform (DFT). The peaks in the DFT domain provide the frequencies of this waveform. However, when the maximal frequency in the waveform is very large, the sampling rate also needs to be very large. Recently, methods to detect a single frequency in an undersampled real-valued waveform have appeared [1 – 4]. A single frequency in a real-valued waveform actually corresponds to two symmetric frequencies in a complex-valued waveform, where their coefficients are complex conjugates of each other. Most recently, general multiple frequency detection in undersampled complex-valued waveforms is studied in [5]. Given L undersampling rates of a complex-valued waveform with K multiple frequencies with $L \geq \eta K$, a range for the detectable frequencies is given in [5], which is the minimum of the least common multiples of all possible η different rates in the L undersampling rates. As mentioned in [5], this range may not be optimal, in particular when some prior information on the multiple frequencies is known. It should be noted that, the larger the range of the detectable multiple frequencies, the better the sampling efficiency, when the multiple sampling rates are given. The main idea in the above method is the implementation of multiple DFTs of the undersampled waveforms and each DFT gives a set of residues of the multiple frequencies modulo the sampling rates. Then, the (generalised) Chinese remainder theorem (CRT) is used in the determination of these multiple frequencies.

In this Letter, we study the case when the multiple frequencies in a complex-valued waveform are assumed in *a priori* to be close to each other within a distance W. Given L undersampling rates, each of which is $> 2W$, we prove that the maximal range for the detectable multiple frequencies is the least common multiple (lcm) of these sampling rates.

Uniqueness of multiple frequency determination: Without loss of generality, the multiple frequencies in a waveform $x(t)$ are assumed $f_1, f_2, \ldots, f_K$ with $f_1 < f_2 < \ldots < f_K$. The complex-valued waveform $x(t)$ is represented by

$$x(t) = \sum_{k=1}^{K} A_k e^{2\pi j f_k t}$$

where A_k, $k = 1, \ldots, K$, are nonzero coefficients. Let m_1 be one

$$x_{m_1}[n] = x\left(\frac{n}{m_1}\right) = \sum_{k=1}^{K} A_k e^{2\pi j f_k n/m_1} \quad n \in \mathbf{Z} \qquad (1)$$

We first consider the single frequency case. In this case $x(t) = A_1 e^{2\pi j f_1 t}$ and $x_{m_1}[n] = A_1 e^{2\pi j r_1 n/m_1}$. Let $f_1 = n_1 m_1 + r_1$, $0 \leq r_1 \leq m_1 - 1$, i.e. $r_1 = f_1 \pmod{m_1}$, then the m_1 point DFT of $x_{m_1}[n]$, $0 \leq n \leq m_1 - 1$, is

$$DFT_{m_1}(x_{m_1}[n]) = A_1 \delta(k - r_1) \qquad 0 \leq k \leq m_1 - 1 \quad (2)$$

This means that the residue $r_1 = f_1 \bmod m_1$ can be detected from the m_1 point DFT of $x_{m_1}[n]$. Now, let $m_2, \ldots, m_L$ be other undersampled rates and $\text{lcm}\{m_1, m_2, \ldots, m_L\}$ be the lcm of the integers in the set $\{m_2, \ldots, m_L\}$. By the CRT we have

Lemma 1: If $f_1 < \text{lcm}\{m_1, m_2, \ldots, m_L\}$, then f_1 can be uniquely determined from the residues $r_l = f_1 \pmod{m_l}$ using the above undersampled DFTs.

We now consider the multiple frequency case. Let $r_{kl} = f_k \bmod m_l$. Then, the m_l point DFT of $x_{m_l}[n]$, $0 \leq n \leq m_l - 1$ is

$$DFT_{m_l}(x_{m_l}[n]) = \sum_{k=1}^{K} A_k \delta(i - r_{kl}) \quad 0 \leq i \leq m_l - 1, \quad 1 \leq l \leq L \qquad (3)$$

From this representation, we see that, without the knowledge of amplitudes (it is usually impossible, for example, if all of the nonzero amplitudes are equal), generally we cannot match the peaks and the residues precisely (see example in Section 3). However, if we know that r_{kl} is the residue of f_k modulo m_l for $l = 1, 2, \ldots, L$, i.e. $r_{kl} = f_k \bmod m_l$ and f_k is upper bounded by $\text{lcm}\{m_1, \ldots, mL\}$, then, by Lemma 1, we can determine the frequencies $f_1, \ldots, f_K$ uniquely. In conclusion, we have the following lemma.

Lemma 2: If $\max\{f_1, \ldots, f_K\} < \text{lcm}\{m_1, \ldots, m_L\}$ and we know the residue r_{kl} of f_k modulo m_l precisely for $1 \leq l \leq L$ and $1 \leq k \leq K$, then we can determine $f_1, \ldots, f_K$ uniquely.

If we only know that the set $\{r_{kl}, k = 1, \ldots K\}_l$ is the set of all residues of f_k, $1 \leq k \leq K$, modulo m_l for $l = 1, 2, \ldots, L$, in general it is impossible to determine which one in the set is the residue of f_k and therefore in general it is impossible to determine the multiple frequencies f_k. It is possible, however, if some prior information about the multiple frequencies f_k is known, which is the goal of the rest of this Letter.

Without loss of generality, we may assume that $f_1 < f_2 < \ldots < f_K$. Let $W = \max_{1 \leq i, j \leq K} |f_i - f_j| = f_K - f_1$. The undersampling rates are chosen $m_1, \ldots, m_L$ such that $\min\{m_1, \ldots, m_L\} > 2W$. Then we have the following lemma.

Lemma 3: Under the above condition on m_l and f_k, $1 \leq l \leq L$, $1 \leq k \leq K$, the residue r_{kl} of f_k modulo m_l is uniquely determined by the set $\{r_{1l}, \ldots, r_{Kl}\}$ for $1 \leq l \leq L$ and $1 \leq k \leq K$.

Proof: Since $\min\{m_1, \ldots, m_L\} > 2W$, $\max\{|f_i - f_j|\} \leq W$ and f_k are all distinct, it is not hard to see that the elements in the set $\{r_{1l}, \ldots, r_{Kl}\}$ are different from each other. Let $\{r_{1l}, \ldots, r_{Kl}\} = \{\alpha_1, \ldots, \alpha_K\}$ with $\alpha_1 < \alpha_2 < \ldots < \alpha_K$. By the following representation

$$\begin{cases} f_1 = r_{1l} + k_1 m_l \\ f_2 = r_{2l} + k_2 m_l \\ \vdots \\ f_K = r_{Kl} + k_K m_l \end{cases} \qquad (4)$$

and the facts that $f_K - f_1 \leq W$ and $m_l > 2W$, we have $k_1 \leq k_2 \leq \ldots \leq k_K$ and $k_K - k_1 = 0$ or 1. If $k_K = k_1$, then $r_{1l} < r_{2l} < \ldots < r_{Kl}$ and $r_{Kl} - r_{1l} < W$. If $k_1 = k_2 = \ldots = k_i$, $k_{i+1} = \ldots = k_K = k_1 + 1$ for some $i \in \{1, 2, \ldots, K-1\}$, then:

$$f_1 = r_{1l} + k_1 m_l, \ldots, f_i = r_{il} + k_1 m_l \qquad (5)$$

and

$$f_{i+1} = r_{(i+1)l} + (k_1 + 1)m_l, \ldots, f_K = r_{Kl} + (k_1 + 1)m_l \quad (6)$$

By eqn. 5, we have $r_{1l} < r_{2l} < \ldots < r_{il}$ and $r_{il} - r_{1l} \leq W$. By eqn. 6, we have $r_{(i+1)l} < r_{(i+2)l} < \ldots < r_{Kl}$ and $r_{Kl} - r_{(i+1)l} \leq W$. Furthermore $r_{1l} - r_{Kl} = m_l - (f_K - f_1) > W$. Hence, we can determine the values of $r_{1l}, \ldots, r_{Kl}$ uniquely according to the following law: if $\alpha_K - \alpha_1 \leq W$, then we have $r_{1l} = \alpha_1, \ldots, r_{Kl} = \alpha_K$. If $\alpha_{i+1} - \alpha_i > W$, for some i, $1 \leq i \leq K-1$ then $r_{1l} = \alpha_{i+1}, \ldots, r_{(K-i)l} = \alpha_K$, $r_{(K-i+1)l} = \alpha_1, \ldots, r_{Kl} = \alpha_i$.

Therefore, in all cases we can determine the residues uniquely.

Combining the above lemmas, we have obtained the following main result.

Theorem 1: Assume complex valued waveform $x(t)$ contains K different frequencies f_k for $1 \le k \le K$. Let m_l, $1 \le l \le L$ be sampling rates in the undersampled versions $x_{m_l}[n]$ of $x(t)$ in eqn. 1 with m_l replaced by m_l, $1 \le l \le L$. Then the K frequencies f_k for $1 \le k \le K$ can be uniquely determined by using the m_l point DFT of $x_{m_l}[n]$ for $1 \le l \le L$ if $\max\{f_1, \ldots, f_K\} < \text{lcm}\{m_1, \ldots, m_L\}$ and $\min\{m_1, \ldots, m_L\} > 2\max_{1 \le i, j \le K}|f_i - f_j|$.

It is clear that the range of f_k, $\text{lcm}\{m_1, \ldots, m_L\}$, given $m_1, \ldots, m_L$ is the maximal one. The difference between the above result and the result in [5] is that the knowledge of $\min\{m_1, \ldots, m_L\} > 2\max_{1 \le i, j \le K}|f_i - f_j|$ is needed in this Letter, while no knowledge is necessary in the result in [5] as mentioned in the introduction.

Multiple frequency determination algorithm: For simplicity, we assume $m_1, \ldots, m_L$ are pairwise coprimes. We assume the conditions in theorem 1 hold. Now, we give the concrete determination algorithm as follows:

Step 1. Sample the waveform $x(t)$ with the sampling rates m_l to obtain $x_{m_l}[n]$ for $1 \le l \le L$.

Step 2. Implement the m_l point DFT of $x_{m_l}[n]$, $0 \le n \le m_l - 1$, to detect the set $S_l = \{\alpha_{1l}, \ldots, \alpha_{Kl}\}$ of K peaks in the DFT domain for $1 \le l \le L$.

Step 3. For each peak set $\{\alpha_1, \ldots, \alpha_K\} = S_l$ with $\alpha_1 < \alpha_2 < \ldots < \alpha_K$, if $\alpha_K - \alpha_1 \le W$, then we have $r_{1l} = \alpha_1, \ldots, r_{Kl} = \alpha_K$. If $\alpha_{i+1} - \alpha_i > W$ and $\alpha_i - \alpha_1 < W$, for some i, $1 \le i \le K-1$ then $r_{1l} = \alpha_{i+1}, \ldots, r_{(K-i)l} = \alpha_K$, $r_{(K-i+1)l} = \alpha_1, \ldots, r_{Kl} = \alpha_i$. Hence we can determine these residues r_{kl} of f_k uniquely.

Step 4. By Step 3 we know:

$$f_k = r_{kl} \pmod{m_l} \qquad k = 1, \ldots, K \quad l = 1, \ldots, L \qquad (7)$$

We define:

$$M = \prod_{l=1}^{L} m_l \quad \text{and} \quad M_l = \frac{M}{m_l} \qquad (8)$$

Since m_l and M_l are coprime, there are solutions N_l of

$$N_l M_l = 1 \pmod{m_l} \qquad (9)$$

With these N_l, the solution f_k is

$$f_k = r_{k1} N_1 M_1 + \cdots + r_{kL} N_L M_L \pmod{M} \qquad (10)$$

The above closed formulas for f_k are the solutions of f_k. For the CRT involved in step 4, see, for example, [7].

Examples: In this Section, we see a simple example. Consider a signal with three frequencies, where their differences are at most 10 Hz, i.e. $W = 10$. We sample this signal with frequencies $m_1 = 27$ Hz, $m_2 = 28$ Hz, $m_3 = 29$ Hz. Hence if the largest frequency of this signal is less than $27 \times 28 \times 29 = 21924$ Hz, by theorem 1, we can uniquely determine these three frequencies. Let $f_1 = 20008$ Hz, $f_2 = 20013$ Hz, and $f_3 = 20017$ Hz.

For the sampling rate $m_1 = 27$ Hz, we obtain the DFT peak values at {1, 6, 10}. For the sampling rate $m_2 = 28$ Hz, we obtain the DFT peak values at {16, 21, 25}. For the sampling rate $m_3 = 29$ Hz, we obtain the DFT peak values at {3, 7, 27}. Then, from step 3 we have

$$r_{11} = 1 \quad r_{21} = 6 \quad r_{31} = 10$$
$$r_{12} = 16 \quad r_{22} = 21 \quad r_{32} = 25$$
$$r_{13} = 27 \quad r_{23} = 3 \quad r_{33} = 7$$

By step 4, the multiple frequencies can be determined.

If we take another sample with rate $m_4 = 31$, then the signal with highest frequency < 679644 Hz can be uniquely determined.

Conclusion: In this Letter, we studied the detection of multiple frequencies in undersampled complex-valued waveforms, where the multiple frequencies are close to each other. Given the undersampling rates, the maximal range of the detectable multiple frequencies was given, that is the lcm of the undersampling rates, when all these rates are larger than twice the maximal distance between the multiple frequencies. The main advantage of undersampling is the hardware cost reduction [6].

Acknowledgments: This work was supported in part by a research initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

21 May 1997
Electronics Letters Online No: 19970891

Guangcai Zhou and Xiang-Gen Xia (*Department of Electrical Engineering, University of Delaware, Newark, DE 19716, USA*)

E-mail: xxia@ee.udel.edu

References

1 PACE, P.E., LEINO, R.E., and STYER, D.: 'Use of the symmetrical number system in resolving single-frequency undersampled aliases', *IEEE Trans. Signal Process.*, 1997, **45**, pp. 1153–1160
2 SANDERSON, R.B., TSUI, J.B.Y., and FREES, N.: 'Reduction of aliasing ambiguities through phase relations', *IEEE Trans. Aerosp. Electron. Syst.*, 1992, **28**, pp. 950–955
3 BROWN, J.L.: 'On the uniform sampling of a sinusoidal signal', *IEEE Trans. Aerosp. Electron. Syst.*, 1988, **24**, pp. 103–106
4 RADER, C.M.: 'Recovery of undersampled periodic waveforms', *IEEE Trans. Acoust. Speech. Signal Process.*, 1977, **25**, pp. 242–249
5 XIA, X.-G.: 'On detection of multiple frequencies in undersampled complex valued waveforms', Preprint, 1997
6 HILL, G.: 'The benefits of undersampling', *Electron. Des.*, 1994, pp. 69–79
7 McCLELLAN, J.H., and RADER, C.M.: 'Number theory in digital signal processing' (Prentice-Hall, Englewood Cliffs, NJ, 1979

Performance of adaptive multisensor decision feedback equaliser for time-varying frequency selective radio channels

S. Buljore, J.F. Diouris and J. Saillard

Indexing terms: Adaptive equalisers, Radio applications

A multisensor decision feedback equaliser based on the minimum mean squared error (MMSE) criterion is studied. The superiority of the performance of the multisensor equaliser is shown by simulation of a whole communication system in which the adaptive equaliser is incorporated. The recursive least squares (RLS) algorithm is used to update the coefficients. From the results obtained for a time-varying urban terrain channel model, the extremely interesting tracking capability of the multisensor equaliser is shown.

Introduction: Frequency selective and time-varying radio channels considerably degrade the performance of OSM-type mobile digital communication systems [1]. In addition to equalisation, the reduction in the variation of the signal to noise ratio (SNR) due to time variation of the radio channels at the receiver using diversity techniques is a desirable asset . This Letter presents a multisensor decision feedback equaliser (DFE) implemented with the recursive least squares algorithm for a time-varying typical urban channel. A Monte Carlo simulation of the digital link incorporating the multisensor equaliser is carried out. The channel model considered is given by the GSM recommendations for a typical urban (TU) environment [2]. The superiority of the performance of the multisensor equaliser to combat intersymbol interference (ISI) and fading is shown. Finally, conclusions on the improved performance and the very interesting tracking ability of the multisensor DFE are drawn.

System and channel model: The scheme of the baseband system simulation is illustrated in Fig. 1. The bit sequences are transmitted using a DQPSK modulation scheme through a square root raised cosine filter. The channel model used is the typical urban channel given in Table 1. The receiver consists of another square root raised cosine filter, an adaptive single/multi-sensor equaliser,

System Identification Using Chirp Signals and Time-Variant Filters in the Joint Time-Frequency Domain

Xiang-Gen Xia, *Member, IEEE*

Reprinted from
IEEE TRANSACTIONS ON SIGNAL PROCESSING
Vol. 45, No. 8, August 1997

System Identification Using Chirp Signals and Time-Variant Filters in the Joint Time-Frequency Domain

Xiang-Gen Xia, *Member, IEEE*

***Abstract*—In this paper, we propose a novel method to identify an unknown linear time invariant (LTI) system in low signal-to-noise ratio (SNR) environment. The method is based on transmitting chirp signals for the transmitter and using linear time-variant filters in the joint time-frequency (TF) domain for the receiver to reduce noise before identification. Due to the TF localization property of chirp signals, a large amount of additive white noise can be reduced, and therefore, SNR before identification can be significantly increased. This, however, cannot be achieved in the conventional methods, where pseudo-random signals are used, and therefore, noise reduction techniques do not apply. Our simulation results indicate that the method proposed in this paper outperforms the conventional methods significantly in low SNR environment. This paper provides a good application of time-frequency analysis and synthesis.**

I. Introduction

THE SYSTEM identification problem is a classical and important problem in signal processing, which has applications in many fields including channel estimation in wireless communications. There have been extensive studies on this problem; see, for example, [2], [3], [28], [31], and [32]. The problem can be stated as

$$y[n] = \sum_k h[n-k]x[k] + v[n] \tag{1.1}$$

where

- $x[k]$ transmitted signal;
- $h[n]$ impulse response of a linear time invariant (LTI) system (or channel);
- $v[n]$ additive noise;
- $y[n]$ received signal.

The problem is to identify the LTI system transfer function $H(\omega)$ of $h[n]$ given the input and the output signals $x[n]$ and $y[n]$.

The conventional method for solving the above problem is the least-squared solution method that is equal to the cross-spectral method in stationary cases, i.e., the system transfer function $H(\omega)$ can be estimated by

$$H(\omega) = \frac{S_{xy}(\omega)}{S_{xx}(\omega)} \tag{1.2}$$

Manuscript received November 11, 1996; revised April 3, 1997. This work was supported in part by an initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377. The associate editor coordinating the review of this paper and approving it for publication was Dr. Yingbo Hua.

The author is with the Department of Electrical Engineering, University of Delaware, Newark, DE 19716 (e-mail: xxia@ee.udel.edu).

Publisher Item Identifier S 1053-587X(97)05791-7.

where $S_{xy}(\omega)$ is the cross-spectrum of $x[n]$ and $y[n]$, and $S_{xx}(\omega)$ is the autospectrum of $x[n]$. When the additive noise $v[n]$ in (1.1) is a zero-mean Gaussian process and statistically independent of the input signal $x[n]$, the estimate in (1.2) is asymptotically unbiased, and its error variance approaches the Cramer–Rao lower bound that is proportional to the variance of the additive noise $v[n]$. Clearly, the performance is limited by this noise variance, or the signal-to-noise ratio (SNR). When this SNR is low, the performance of the estimate in (1.2) is poor. Since the autospectrum of the input signal $x[n]$ is in the denominator in the estimate (1.2), the input signal is, in general, chosen as a pseudo-random signal with flat spectrum [4]. With these kinds of input signals, noise reduction techniques before system identification do not apply. As a matter of fact, any traditional noise reduction technique, such as any Fourier transform technique, does not perform well for wideband signals. This implies that it is not possible to increase the SNR or the performance of the estimate (1.2) by transmitting a pseudo random signal and using the conventional Fourier noise reduction techniques. Several questions arise here:

i) Can we transmit other wideband signals, such as chirp signals, instead of pseudo random signals?
ii) If so, can we take the advantage of these wideband signals and reduce the noise $v[n]$ in (1.1)?
iii) If so, can we improve the performance of the estimate (1.2) after denoising?

The aim of this paper is to positively answer these questions. The main idea is the following. Chirp-type signals are transmitted, which have wideband characteristics in the frequency domain but concentrate in the joint time-frequency domain. Chirp-type signals are used quite often, such as in radar and in FM in communications systems. The TF concentration property usually holds after an LTI system (this will be seen later). Since a joint TF distribution usually spreads noises and localizes signals, in particular chirps, the receiver may use a TF analysis technique (see, for example, [5]–[27]) to map the received signal $y[n]$ from the time domain into the joint time-frequency domain. In this way, the SNR can be significantly increased in the joint TF domain, and the receiver may be able to see patterns in the joint TF plane and therefore reduce the noise by filtering in the joint TF domain. This filtering is basically a *time-variant* filtering. We use this name in the rest of this paper. The model (1.1) after a time-variant filter

becomes

$$\tilde{y}[n] = \sum_k h[n-k]x[k] + \tilde{v}[n] \tag{1.3}$$

where $\tilde{v}[n]$ is the new noise after the filtering.

The time-variant filter used in this paper is based on the discrete Gabor transforms, which was studied in [5]–[7]. For chirp-type signals, about 13 dB SNR is increased consistently with this filter in [6]. When the original SNR in (1.1) is not too low, say, for example, above -1 dB, the new SNR in (1.3) may reach a significant high level so that the estimate of $H(\omega)$ from $\tilde{y}[n]$ and $x[n]$ is accurate enough for many applications. In this paper, both denoising with several mask design methods and system identification simulations are performed. These simulations show that a much better performance over the conventional method can be achieved.

It should be pointed out that the optimal training signal design for dynamic system identification has a long history dating back over 20 years. The design methods are traditionally based on the minimization of the Cramer–Rao bound for the system parameter estimation in either the time or the frequency domain (see, for example, [28]–[32]) but not in the joint TF domain. The aim of this paper is, however, not focused on the optimal training signal design, although it is a very interesting topic. Denoising before identification using nonredundant discrete wavelet transform was studied in [33] for chemical process control applications.

This paper is organized as follows. In Section II, we briefly review discrete Gabor transforms and the iterative time-variant filtering studied in [5]–[7]. In Section III, we use the time-variant filter studied in Section II to reduce additive white Gaussian noise for a received signal. The filtering problem in this paper has its own characteristics due to the fact that the transmitter and the receiver know the transmitted chirp signal $x[n]$, and therefore, its TF information is known *a priori*. This TF information can be used in designing a mask in the time-variant filtering. In Section IV, we utilize the conventional system identification method, i.e., the cross-spectral method (1.2), after the denoising in Section III. In Section V, we conclude this paper by addressing some possibilities for further improvements.

II. Discrete Gabor Transform and Time-Variant Filtering

There have been many TF analysis techniques, such as Wigner–Ville distributions in the Cohen's class, spectrogram (short-time Fourier transform or Gabor transform or DFT filterbanks), and scalogram (wavelets) (see, for example, [5], [23]–[27]). Some of them, such as bilinear TF distributions, have high resolution but have crossterms for multicomponent signals. Some of them, such as linear techniques (for example, Gabor transforms and wavelet transforms), do not have crossterms for multicomponent signals but may not have very high resolutions. Since, in this paper, we deal with a linear combination (or a linear system) of various chirp signals, it is important for a TF analysis technique not to have crossterms while it should also have a good resolution. This leads us to consider Gabor transforms. In this section, we first review the discrete Gabor transforms (DGT).

Since oversampled DGT is more robust for noise, it is usually used in noise reduction applications. However, a disadvantage for oversampled DGT is that it is not an onto mapping. In other words, not every signal $S[k,l]$ in the DGT transform domain corresponds to a time domain signal $s[n]$ so that the DGT of $s[n]$ is exactly equal to $S[k,l]$. This causes problems in filtering in the DGT transform domain, which is that the filtered signal in the DGT transform domain may not correspond to any time domain signal as shown in Fig. 1. An intuitive solution for this problem is to take the least-squared error (LSE) solution in the time domain (see, for example, [8]–[13]). The LSE, however, usually does not have a desired TF characteristics in the DGT transform domain. When a signal is very long, the computational load for the LSE solution is significantly high because of the inverse matrix computation. Based on these observations, an iterative algorithm was proposed in [5]–[7]. Conditions on the convergence, properties of the limit signals, and the relationship between the LSE solutions and solutions from the iterative algorithms were obtained in [6] and [7], where a significant improvement over the LSE solution was also shown. The second part of this section is to briefly review some of these results.

A. Discrete Gabor Transform

We first review some basics on the DGT, which is necessary for this paper. For more about the discrete short-time Fourier transform, see [14], for more about DFT filterbanks, see [15], and for more about the DGT, see, for example, [16]–[22]. Let a signal $s[k]$, a synthesis window function $h[n]$, and an analysis window function $\gamma[n]$ be all periodic with same period L. Then

$$s[k] = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} C_{m,n}h_{m,n}[k] \tag{2.1}$$

$$C_{m,n} = \sum_{k=0}^{L-1} s[k]\gamma^*_{m,n}[k] \tag{2.2}$$

$$h_{m,n}[k] = h[k-m\Delta M]W_L^{n\Delta Nk} \tag{2.3}$$

$$\gamma_{m,n}[k] = \gamma[k-m\Delta M]W_L^{n\Delta Nk} \tag{2.4}$$

and $W_L = \exp(j2\pi/L)$, $j=\sqrt{-1}$. The coefficients $C_{m,n}$ are called the DGT of the signal $s[k]$, and the representation (2.1) is called the *inverse DGT* (IDGT) of the coefficients $C_{m,n}$. One condition on the analysis and synthesis window functions $\gamma[k]$ and $h[k]$ obtained by Wexler and Raz is the identity[1]

$$\sum_{k=0}^{L-1} h[k+mN]W_L^{-nMk}\gamma^*[k] = \delta[m]\delta[n], \quad 0\le m\le \Delta N-1, \quad 0\le n\le \Delta M-1 \tag{2.5}$$

[1]If we take the inverse discrete Fourier transform with respect to the parameter n at the both sides, the system (2.5) is the same as the one obtained in [14] when all convolutions are considered to be cyclic convolutions for finite length signals in [14].

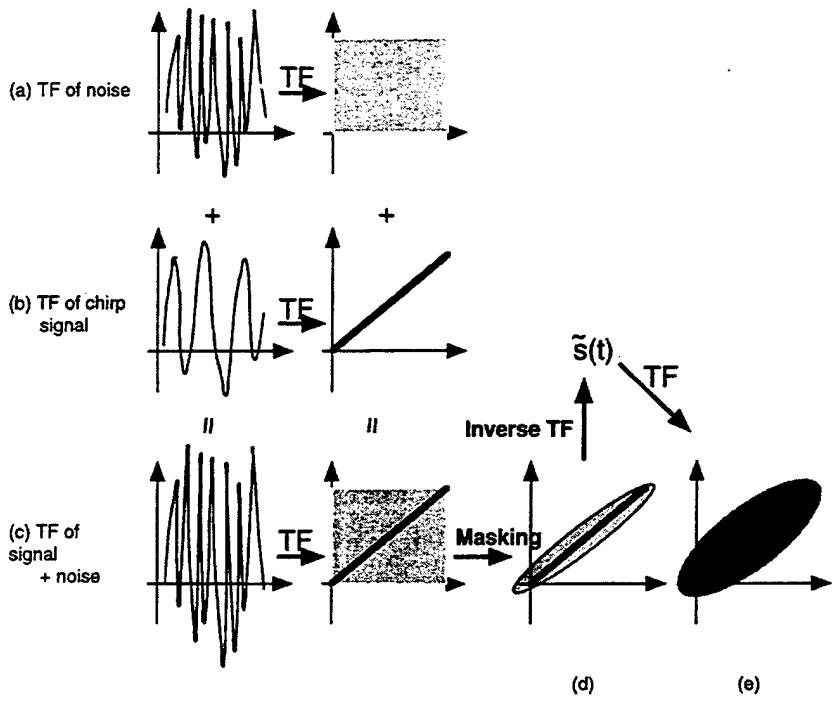

Fig. 1. TF transform illustration.

where ΔM and ΔN are the time and the frequency sampling interval lengthes, and M and N are the numbers of sampling points in the time and the frequency domains, respectively, $M \cdot \Delta M = N \cdot \Delta N = L$, $MN \geq L$ (or $\Delta M \Delta N \leq L$). The critical sampling case is when $M \cdot N = \Delta M \cdot \Delta N = L$. The condition (2.5) on window functions h and γ can be rewritten in matrix form as

$$H_{p\times L}\gamma^*_{L\times 1} = \mu_{p\times 1} \tag{2.6}$$

where the subscript $m \times n$ means the m by n matrix $p = \Delta M \cdot \Delta N, \gamma_{L\times 1} = (\gamma[0], \gamma[1], \cdots, \gamma[L-1])^T$, and $\mu_{p\times 1} = (1, 0, \cdots, 0)^T$ and the element at the $(m\Delta M + n)$th row and the kth column in the matrix $H_{p\times L}$ is

$$h[k + mN]W_L^{-nMk}, \qquad 0 \leq m \leq \Delta N - 1$$
$$0 \leq n \leq \Delta M - 1, \quad 0 \leq k \leq L - 1.$$

In the critical sampling case and when $H_{p\times L}$ has full rank, there is a unique solution for the analysis window function $\gamma[n]$. In the oversampling case and when $H_{p\times L}$ has full rank, there are infinite many solutions for the system (2.5). Among them, the minimum norm solution was given in [17]

$$\gamma^*_{L\times 1} = H^\dagger_{p\times L}(H_{p\times L}H^\dagger_{p\times L})^{-1}\mu_{p\times 1} \tag{2.7}$$

where † means the complex conjugate transpose. It was proved in [18]–[20] that the above minimum norm solution is also the most orthogonal-like solution, i.e., (a more general form was given in [22])

$$\|\gamma_{L\times 1} - h_{L\times 1}\| = \min_{\hat{\gamma}_{L\times 1}: H_{p\times L}\hat{\gamma}^*_{L\times 1} = \mu_{p\times 1}} \|\hat{\gamma}_{L\times 1} - h_{L\times 1}\|. \tag{2.8}$$

The DGT and IDGT can be also represented in matrix forms. Let

$$C = (C_{0,0}, C_{0,1}, \cdots, C_{M-1,N-1})^T$$
$$s = (s[0], s[1], \cdots, s[L-1])^T.$$

The DGT can be represented by the $MN \times L$ matrix $G_{MN\times L}$ with its $(mN + n)$th row and kth column element

$$\gamma^*_{m,n}[k] = \gamma^*[k - m\Delta M]W_L^{-n\Delta Nk}, \qquad 0 \leq m \leq M - 1$$
$$0 \leq n \leq N - 1, \quad 0 \leq k \leq L - 1.$$

The IDGT can be represented by the $L \times MN$ matrix $H_{L\times MN}$ with its kth row and $(mN + n)$th column element

$$h_{m,n}[k] = h[k - m\Delta M]W_L^{n\Delta Nk}, \qquad 0 \leq m \leq M - 1$$
$$0 \leq n \leq N - 1, \quad 0 \leq k \leq L - 1.$$

Thus

$$C = G_{MN\times L}s \quad \text{and} \quad s = H_{L\times MN}C. \tag{2.9}$$

The condition (2.5) implies that

$$H_{L\times MN}G_{MN\times L} = I_{L\times L} \tag{2.10}$$

where $I_{L\times L}$ is the $L \times L$ identity matrix.

B. Iterative Time-Variant Filtering Algorithm

We next want to briefly review the iterative time-variant filtering algorithm proposed in [5]–[7]. This algorithm is used later in the denoising for the system identification problem.

The oversampling of the DGT adds redundancy, which is usually preferred for noise reduction applications. From (2.1)–(2.5), (2.9), and (2.10), one can see that an L-dimensional signal s is transformed into an MN-dimensional

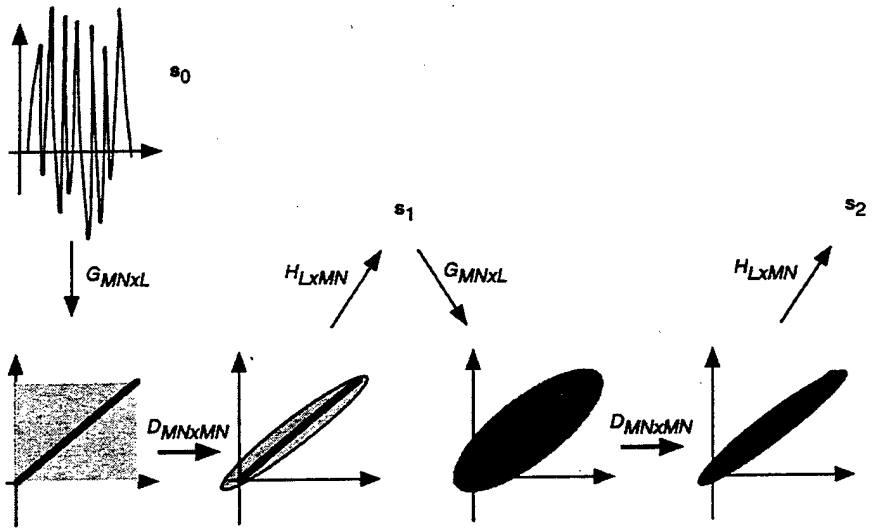

Fig. 2. Iterative time-varying filtering algorithm.

signal C, and MN is greater than L due to the oversampling. Therefore, only a small set of MN-dimensional signals in the TF plane have their corresponding time waveforms with length L. Let $D_{MN\times MN}$ denote the mask transform, specifically, a diagonal matrix with diagonal elements either 0 or 1. Let $\boldsymbol{s}$ be a signal with length L in the time domain. The first step in the time-variant filtering is to mask the TF transform of $\boldsymbol{s}$

$$C_1 = D_{MN\times MN}G_{MN\times L}\boldsymbol{s}$$

where $D_{MN\times MN}$ masks a desired domain in the TF plane. Since the DGT $G_{MN\times L}$ is a redundant transformation, the IDGT of $C_1, H_{L\times MN}C_1$ may not fall in the mask. In other words, in general

$$G_{MN\times L}H_{L\times MN}C_1 \neq D_{MN\times MN}G_{MN\times L}H_{L\times MN}C_1 \tag{2.11}$$

where $MN > L$, which is illustrated in Fig. 1(e). Notice that in the critical sampling case, i.e., $MN = L$, the inequality (2.11) becomes an equality. An intuitive method to reduce the difference between the right- and the left-hand sides of (2.11) is to mask the right-hand side of (2.11) again and repeat the procedure, which leads to the iterative algorithm

$$\boldsymbol{s}_0 = \boldsymbol{s} \tag{2.12}$$

$$C_{l+1} = D_{MN\times MN}G_{MN\times L}\boldsymbol{s}_l \tag{2.13}$$

$$S_{l+1} = H_{L\times MN}C_{l+1}, \qquad l = 0, 1, 2, \cdots. \tag{2.14}$$

The above iterative algorithm is illustrated in Fig. 2.

Before going to the convergence, let us see what the LSE is. Based on the definition, the LSE solution is the $L \times 1$ vector $\overline{\boldsymbol{x}}$ that minimizes

$$\|G_{MN\times L}\overline{\boldsymbol{x}} - D_{MN\times MN}G_{MN\times L}\boldsymbol{s}\| = \min_{\boldsymbol{x}} \|G_{MN\times L}\boldsymbol{x} - D_{MN\times MN}G_{MN\times L}\boldsymbol{s}\|. \tag{2.15}$$

Then

$$\overline{\boldsymbol{x}} = (G^{\dagger}_{MN\times L}G_{MN\times L})^{-1}G^{\dagger}_{MN\times L}D_{MN\times MN}G_{MN\times L}\boldsymbol{s}. \tag{2.16}$$

Clearly, when the signal length L is large, the inverse matrix computation is expensive. Although the error in (2.15) is minimized, the DGT of the least-squared solution $\overline{\boldsymbol{x}}$ may not fall in the mask $D_{MN\times MN}$: $G_{MN\times L}\overline{\boldsymbol{x}} \neq D_{MN\times MN}G_{MN\times L}\overline{\boldsymbol{x}}$ when $MN > L$.

The complexity for the iterative algorithm (2.12)–(2.14) is, however, low, which does not need to compute inverses of large size matrices. By considering the DGT and IDGT in (2.1)–(2.4), the computational complexity in (2.12)–(2.14) is proportional to the signal length multiplied by the window length, i.e., LL_W. Notice that the complexity of directly computing the inverse matrices in (2.16) is proportional to L^3. Therefore, when the length of window functions h and γ is much shorter than the length of the signal $\boldsymbol{s}$, the computational complexity in the iterative algorithm (2.12)–(2.14) is much lower than the one for the least-squared solution in (2.16).

We next want to list several related results on the above iterative algorithm obtained in [6] and [7], such as the convergence, the properties of the limit signals, and the relationship between this algorithm and the LSE solution. These results are based on the condition on the window functions h and γ obtained in [6] and [7]:

$$\sum_{l=0}^{\Delta N-1} \gamma^*[lN+k]h[lN+k+m\Delta M] = \sum_{l=0}^{\Delta N-1} h^*[lN+k]\gamma[lN+k+m\Delta M] \tag{2.17}$$

for $k = 0, 1, \cdots, N-1$ and $m = 0, 1, \cdots, M-1$.

Theorem 1: When the synthesis and the analysis window functions $h[n]$ and $\gamma[n]$ satisfy condition (2.17), the iterative algorithm (2.12)–(2.14) converges.

There are two trivial cases where (2.17) holds. The first case is the orthogonal case $h[n] = \gamma[n]$ for all integer n. The second case is the critical sampling case $\Delta M = N$. Notice that the continuous Gabor transform is never orthogonal unless the window functions are badly localized in the frequency domain. This, however, is not the case for the DGT. The

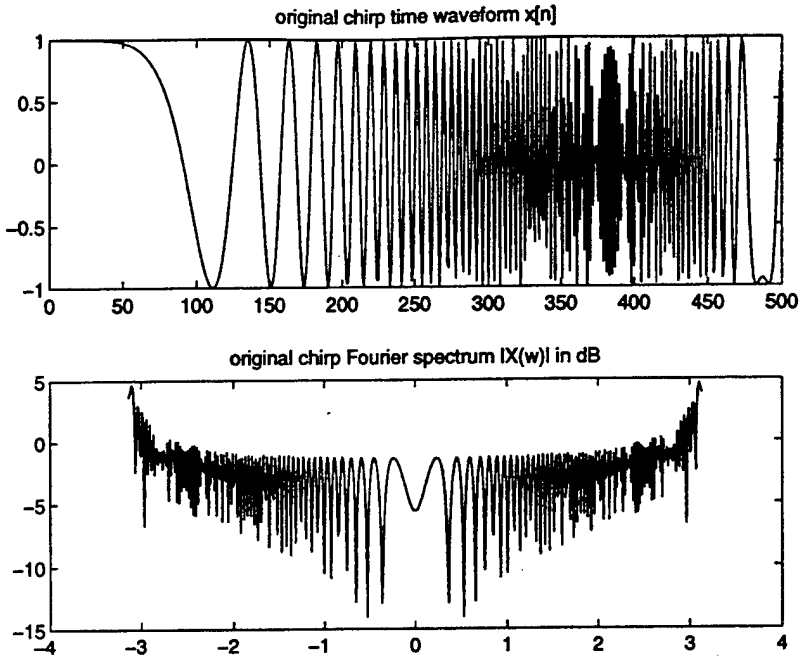

Fig. 3. Transmitted signal $x[n]$ and its Fourier spectrum $X(\omega)$.

most orthogonal-like solution was studied by Qian *et al.* in [18]–[20]. They showed that it is possible to have the analysis window function γ very close to the synthesis window function h when h is truncated Gaussian. The error between h and γ is less than 2×10^{-6} (see Fig. 4) while they are of unit energy, and therefore, the error is negligible. It was shown in [6] that the performance of the iterative algorithm strongly depends on (2.17). When this condition does not satisfy, the iterative algorithm may not converge.

Theorem 2: Under (2.17), the DGT of the limit $\bar{\boldsymbol{s}}$ of the iterative algorithm (2.12)–(2.14) falls in the mask $D_{MN\times MN}$, i.e.

$$G_{MN\times L}\bar{\boldsymbol{s}} = D_{MN\times MN}G_{MN\times L}\bar{\boldsymbol{s}}. \tag{2.18}$$

The above results say that as long as (2.17) on the analysis and synthesis window functions is satisfied, the iterative algorithm converges, and the limit signal has the desired TF characteristics, i.e., its DGT falls in the desired mask. One might ask whether it violates the known fact that an image of a TF transform of a signal in the TF plane cannot be compactly supported. This is because a signal cannot be time- and bandlimited simultaneously. To answer this question, we first need to know that the above known fact is true for continuous TF transforms. Moreover, the proof of the fact is based on the marginal properties of TF transforms. It may not be true for discrete TF transforms. In other words, discrete TF transforms may have compact support [5].

Theorem 3: Under (2.17), the first iteration $\boldsymbol{s}_1$ of the iterative algorithm (2.12)–(2.14) is equal to the least-squared solution in (2.16), i.e., $\boldsymbol{s}_1 = \bar{\boldsymbol{x}}$.

With this result, one will see later that the iterative algorithm (2.12)–(2.14) improves the least-squared solution when the number of iterations increases, and meanwhile, one does not need to compute the inverse matrix in (2.16).

III. Denoising for Received Signals Through a Noisy Channel

In this section, we want to do noise reduction with the time-variant filter studied in Section II for received signals in a noisy channel.

A. Some Parameters

The signal length is randomly chosen as 500. The signal $x[n]$ for the transmitter is

$$x[n] = \cos\left(\left[\frac{n+15}{150}\right]^4\right), \qquad n = 0, 1, \cdots, 499. \tag{3.1}$$

The waveform and its Fourier transform $X(\omega)$ of the above signal $x[n]$ are shown in Fig. 3. Notice that since the Fourier power spectrum $|X(\omega)|^2$ will be used in the denominator in the system identification, it should be as far away from zero as possible. Since the noise-reduction performance of the time-variant filtering in Section II depends on the localization of the signal in the TF plane, the transmitted signal $x[n]$ should be as concentrated in the joint time and frequency domain as possible. The synthesis and analysis window functions used in this paper are shown in Fig. 4, where their lengthes are 256. The synthesis window function is just the Gaussian function and its analysis window function is the most orthogonal-like solution given in (2.7). Their difference and the difference between the left-hand side and the right-hand side of (2.17), i.e., the condition error, are also shown in Fig. 4. One can see that they almost satisfy (2.17). The time sampling interval length $\Delta M = 16$ and the frequency sampling interval length $\Delta N = 2$ in the discrete Gabor transform and its inverse in Section II. These parameters are used throughout the rest of this paper. The DGT of $x[n]$ is shown in Fig. 5. The tail part

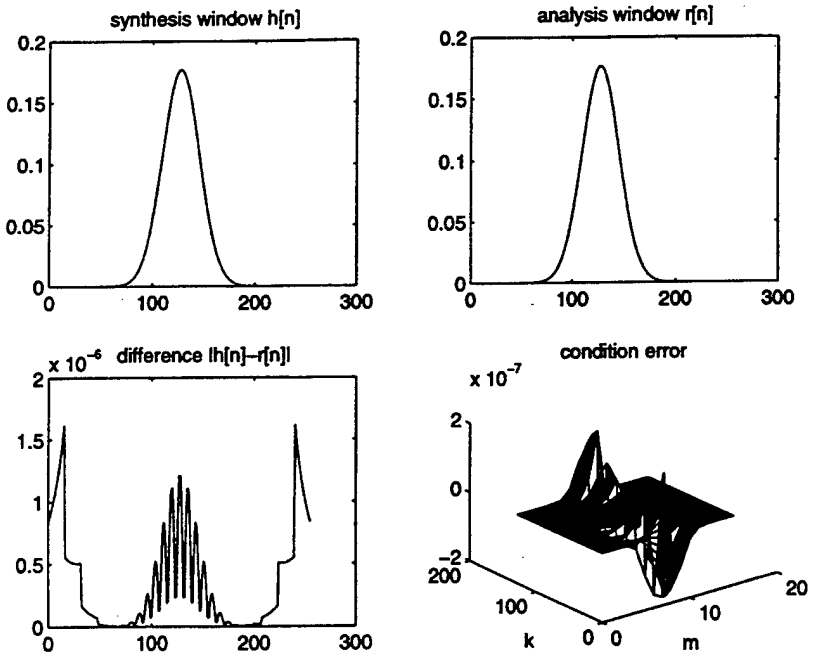

Fig. 4. Synthesis and analysis window functions and the condition (2.17) test.

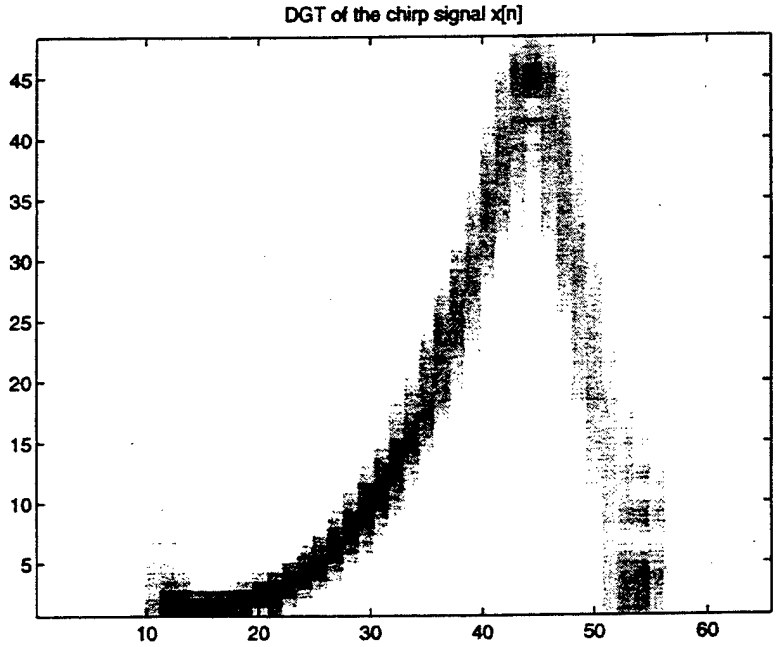

Fig. 5. Discrete Gabor transform of signal $x[n]$.

of the DGT in Fig. 5 is because of the discrete calculation aliasing.

In this paper, we use 20-tap LTI systems in our numerical examples, where the number 20 is just randomly chosen. The channel model is

$$y[n] = \sum_{k=0}^{N-1} h[k]x[n-k] + v[n] \tag{3.2}$$

where $N = 20$ in the following numerical examples, $v[n]$ is an additive white Gaussian noise and independent of the signal $x[n]$, and

$$s[n] = \sum_{k=0}^{N-1} h[k]x[n-k] \tag{3.3}$$

is considered to be the signal, $x[n]$ is the transmitted signal as in (3.1), $y[n]$ is the received signal, and $h[n]$ is an LTI system (or channel) impulse response. The original SNR for

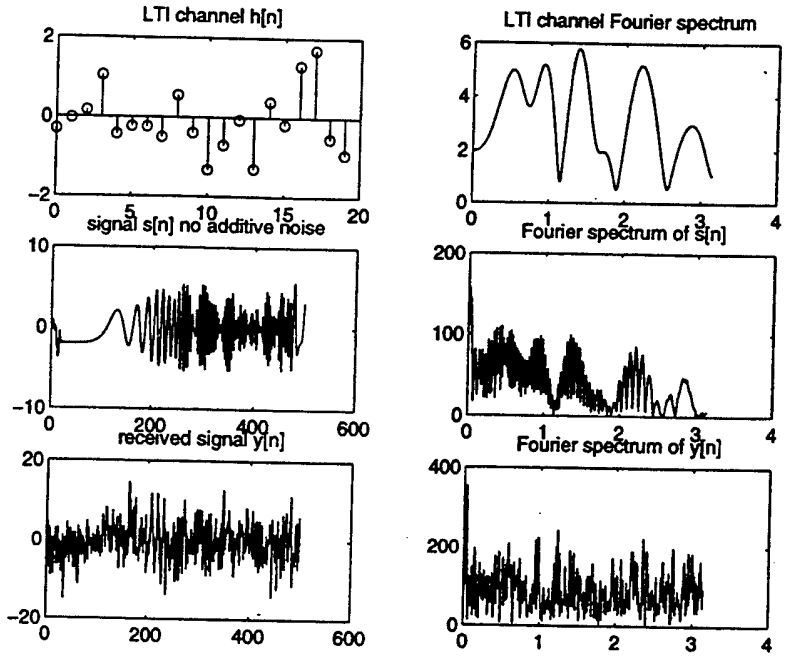

Fig. 6. Example of LTI channel $h[n]$, signal $s[n]$, and received signal $y[n]$ and their Fourier spectrum, where the SNR $= -4.5$ dB for the additive white Gaussian noise.

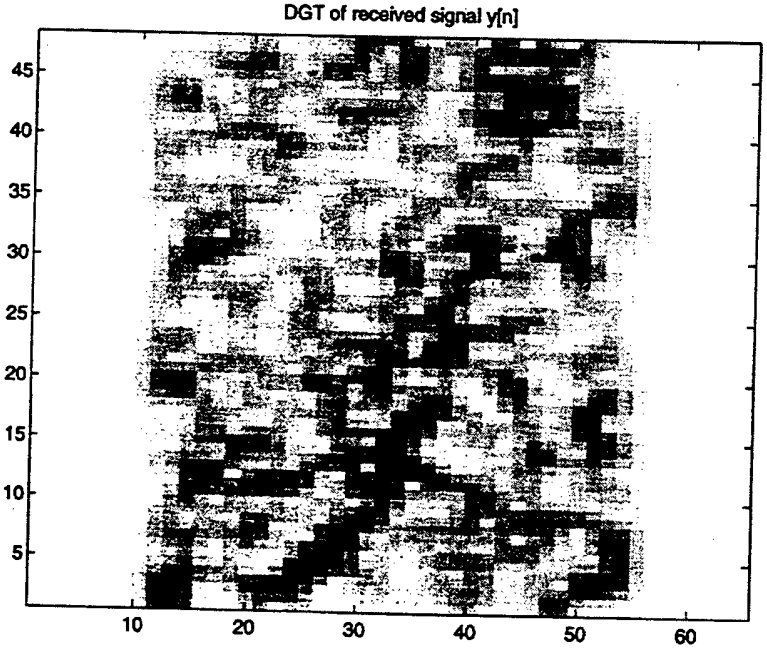

Fig. 7. Discrete Gabor transform of the received signal $y[n]$ in Fig. 6 with SNR $= -4.5$ dB.

the received signal is calculated by

$$10\log_{10}\left(\frac{\sum_{n=0}^{499}|s[n]|^2}{\sum_{n=0}^{499}|v[n]|^2}\right).$$

In the following, we randomly generate the channel $h[n]$. As an example, a channel Fourier spectrum and received signal time waveform $y[n]$ with SNR $= -4.5$ dB and the signal $s[n]$ without noise and their Fourier spectrum are shown in Fig. 6. The DGT of the received signal $y[n]$ with -4.5 dB SNR is shown in Fig. 7. In Fig. 7, one is still able to see the chirp pattern in the joint time and frequency plane, although it is impossible in the time or the frequency domain alone in Fig. 6.

B. Mask Design

The pattern in the DGT domain of the above signal $s[n]$ in (3.3) is similar to the one for the signal $x[n]$ in Fig. 5. This is

not only true for this particular example but is also the case for our numerous examples. The reason is due to the following analytic argument.

Assume the chirp signal $x[n] = \exp(jcn^r)$ for some constants $r \geq 2$ and $c \neq 0$. Then

$$\begin{aligned} s[n] &= \sum_k h[k]x[n-k] \\ &= \sum_k h[k]\exp(jc(n-k)^r) \\ &= \exp(jcn^r)\sum_k h[k]\exp\left(jc\sum_{l=0}^{r-1} c_l n^l k^{r-l}\right) \\ &= x[n]\sum_k h[k]\exp\left(jc\sum_{l=0}^{r-1} c_l n^l k^{r-l}\right) \end{aligned}$$

which is dominated by the original chirp $x[n]$ for finite tap LTI systems $h[k]$. It is because that the highest chirp order of $s[n], r$, and the corresponding chirp rate are the same as those of $x[n]$, whereas the chirp order for the above multiplier of $x[n]$ in $s[n]$

$$\sum_k h[k]\exp\left(jc\sum_{l=0}^{r-1} c_l n^l k^{r-l}\right)$$

is only $r-1$. As a special case, when $r=2$

$$s[n] = x[n]G(2cn)$$

where $G(\omega)$ is the Fourier transform of the signal $h[n]x[n]$

$$G(\omega) = \sum_k h[k]x[k]\exp(-j2cnk).$$

When the channel $h[n]$ has only a finite tap, the function $G(\omega)$ is usually a smooth signal.

Since the transmitted signal $x[n]$ is known to both transmitter and the receiver, by the above property, its pattern in the DGT domain may help in designing a mask in the DGT domain for filtering noise. This is exactly the motivation for the following design method of a mask $D_{MN\times MN}$ in the iterative time-variant algorithm (2.12)–(2.14). The subscript $MN \times MN$ of the mask $D_{MN\times MN}$ will be dropped from now on without causing confusion in understanding.

1) Mask Design Procedure:

Step 1) Implement the DGT $C_{m,n}$ of the transmitted signal $x[k]$.

Step 2) Threshold the DGT coefficients $C_{m,n}$ and have a mask D_x from $C_{m,n}$

$$D_x(m,n) = \begin{cases} 1, & \text{if } |C(m,n)| > t_0 \\ 0, & \text{otherwise} \end{cases}$$

where t_0 is a predesigned positive number that is called *thresholding constant.*

Step 3) Implement Steps 1 and 2 for the received signal $y[k]$, and design a mask D_y with thresholding constant t_1 from the DGT coefficients of $y[n]$ with another predesigned constant $t_1 > 0$.

Step 4) The final mask is the product of D_x and D_y: $D = D_xD_y$.

Since the DGT of the signal $x[n]$ usually dominates the DGT of the signal $s[n]$, the pattern in the DGT domain of the signal $s[n]$ is usually in a close neighborhood of the pattern in the DGT domain of $x[n]$. Therefore, the mask D_x is usually designed so that it covers a relatively large area, i.e., the thresholding constant t_0 in Step 2 is usually chosen not too large. Since the received signal $y[n]$ is from a noisy channel, the resolution of its DGT pattern may be reduced, and therefore, the thresholding constant t_1 in Step 3 is usually chosen to be not too small. Otherwise, the mask D_y will cover too much unwanted area. Let us see an example. The mask D_x from $x[n]$, the mask D_y from $y[n]$, their product $D = D_xD_y$, and the mask D_s from the true signal $s[n]$ are shown in Fig. 8, respectively. The SNR in this case is SNR $= -1.4$ dB. The thresholding constants in Steps 1–3 are $t_0 = 0.12$ and $t_1 = 0.15 \cdot \max(\text{DGT}(y))$. It should be pointed out that the above mask design procedure may be improved by using more sophisticated designs. Possible improvements are

i) to find the optimal thresholding constants t_0 and t_1 by training a large number of signals and systems;

ii) to use more sophisticated statistical detection method in the DGT domain for the received signal $y[n]$ instead of a simple thresholding in Step 3;

iii) to smooth the mask $D = D_xD_y$ since the true mask D_s is usually smooth due to the nature of a chirp signal, but D_y from the noisy signal $y[n]$ may not be smooth. Some morphological operations, such as dilation, may be used to smooth the mask D.

Another observation from our various numerical examples is that the mask D_x is the mean of the true mask D_s in terms of different LTI systems $h[n]$.

C. Denoising Experiments

In this subsection, we want to implement the time-variant filtering algorithm in Section II with three masking techniques: using the mask $D = D_x$ from the transmitted signal, using the mask $D = D_yD_x$ as designed by Steps 1–4, using the true mask $D = D_s$. We run 100 tests in terms of different LTI systems $h[n]$ (randomly generated) and different additive white Gaussian noises $v[n]$ for each masking method and take their mean SNR. Nine iterative steps are used in the iterative algorithm (2.12)–(2.14). Fig. 9 shows the curves of the mean SNR versus iterative steps for the three masking methods.

First, we analyze the time-variant filter (2.12)–(2.14) with the mask $D = D_x$. From Fig. 9, the SNR drops after the second iteration. This is because the mask we used is $D = D_x$, which matches the transmitted signal $x[n]$ and not $s[n]$. Although there is a similarity (see Fig. 8) in the TF plane between the DGT of $x[n]$ and the DGT of $s[n]$, they are not equal. The similarity is exactly the reason why the SNR increases significantly in the first and the second iteration step. The difference between $x[n]$ and $s[n]$ causes the SNR to drop after the second iteration. Notice that the mask $D = D_x$ is known to the receiver, and it is a good candidate in the time-variant filtering if the iterative algorithm stops at the second iteration step.

We now analyze the performance of the mask $D = D_xD_y$. This mask rejects a lesser portion of the noise outside D_s than D_x alone does, when the first thresholding constant t_0 for D_x

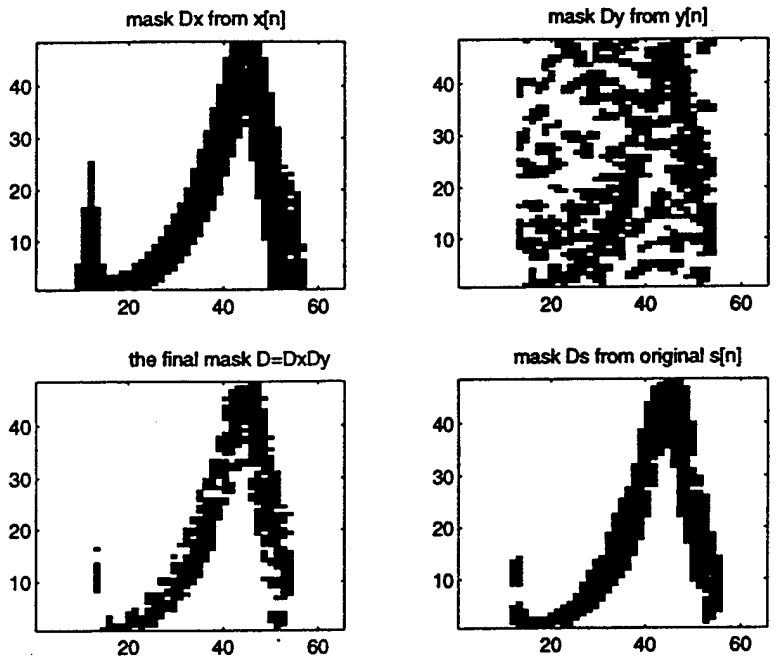

Fig. 8. Example of masks D_x from $x[n]$, D_y from $y[n]$, the final mask $D = D_xD_y$, and the true mask D_s from $s[n]$, where the SNR $= -1.4$ dB.

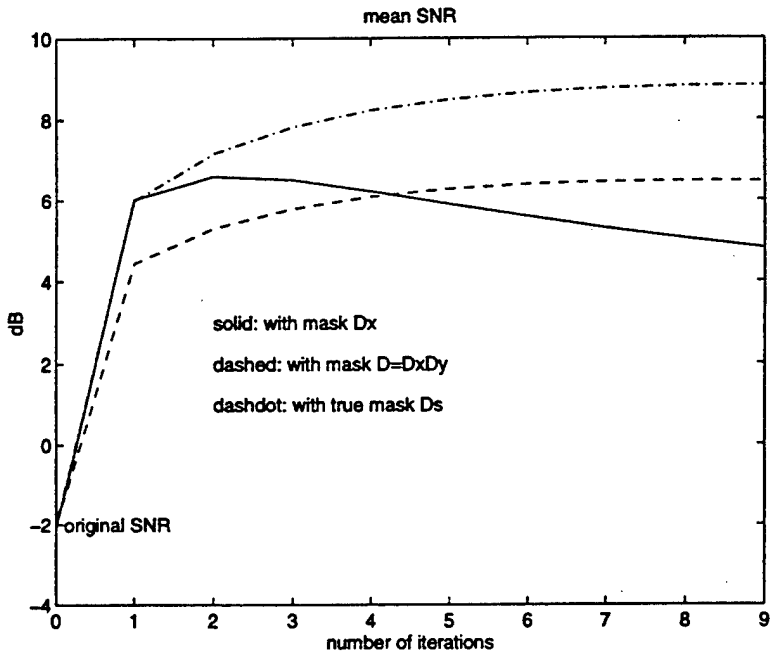

Fig. 9. Mean SNR curves of the iterative time-variant filtering with the following masks: $D = D_x$, $D = D_xD_y$, and $D = D_s$.

in Step 2 is less than the one in designing D_x alone. The reason why this t_0 for D should not be large is for the conservation because the mask D_x is multiplied by D_y in designing D. It, however, happens because the beginning SNR's are not as high as the ones in the time-variant filtering with the mask D_x, which is shown by the solid line in Fig. 9. Since, in general, $D = D_xD_y$ covers relatively more signal information than D_x alone does, the SNR increases when the iteration number increases.

The third masking $D = D_s$ method is the ideal case. With this ideal mask, about an 11 dB SNR increase with the iterative time-variant filtering over the original SNR is achieved consistently. Notice that by Theorem 3, the first iteration is equal to the conventional least squared solution. The iterative time-variant filtering outperforms the least squared solution by about 3 dB.

To improve the performance of the iterative time-variant filtering, what one can do further is to use more sophisticated methods to detect D_x and D_y, in particular D_y, so that their product $D = D_xD_y$ is as close to D_s as possible. Besides what has been mentioned in the previous subsections, directly minimizing the difference between $D = D_xD_y$ and

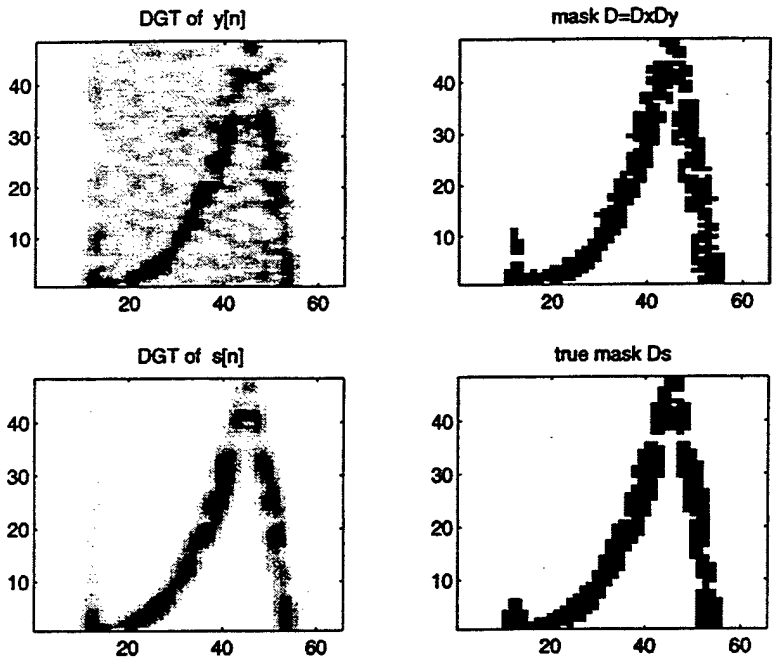

Fig. 10. DGT of $y[n]$ with noise and $s[n]$ without noise and their corresponding masks (original SNR =2.7 dB).

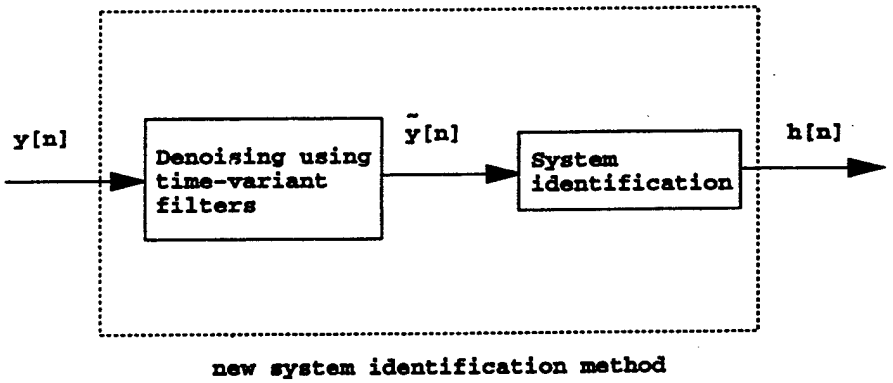

Fig. 11. New system identification method.

D_s with training signals is another potential approach. When the original SNR is not too low, the chirp pattern of $s[n]$ can usually be seen clearly in the DGT domain of the received signal $y[n]$. An example is shown in Fig. 10, where the original SNR = 2.7 dB.

IV. System Identification

In this section, we first use the iterative time-variant filter (2.12)–(2.14) developed in the previous sections to reduce the additive white Gaussian noise $v[n]$ from the received signal $y[n]$. In the iterative time-variant filter, for calculation simplicity, we choose the first masking method studied in Section III-C, i.e., the mask $D = D_x$, for all calculations in this section. With this mask, two iterations are used in the time-variant filter in Section II-B. We then implement the conventional system identification method, as shown in Fig. 11.

The conventional system identification method used here is the cross-spectral method

$$H_{\text{new}}(\omega) = \frac{S_{\tilde{y}x}(\omega)}{S_{xx}(\omega)} \qquad (4.1)$$

where $x[n]$ is the chirp signal defined in (3.1). It is compared with the conventional method without denoising, i.e.,

$$H_{\text{old}_1}(\omega) = \frac{S_{yx}(\omega)}{S_{xx}(\omega)} \qquad (4.2)$$

where $x[n]$ is also the chirp signal. Since the system identification performance usually depends on the signal $x[n]$ transmitted, one might say that it is not fair to compare them using the chirp signal that is preferred here for denoising but might not be preferred for other methods. For this reason, we also compare our new method with the conventional method

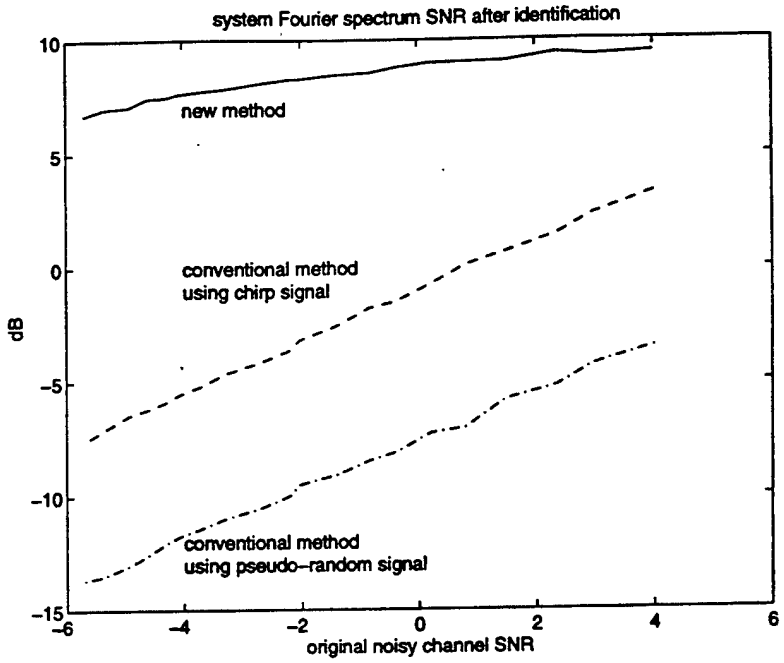

Fig. 12. Comparison of system identification methods. The conventional method using chirp signals; the conventional method using pseudo-random signals; new method using chirp signals, and time-variant filtering.

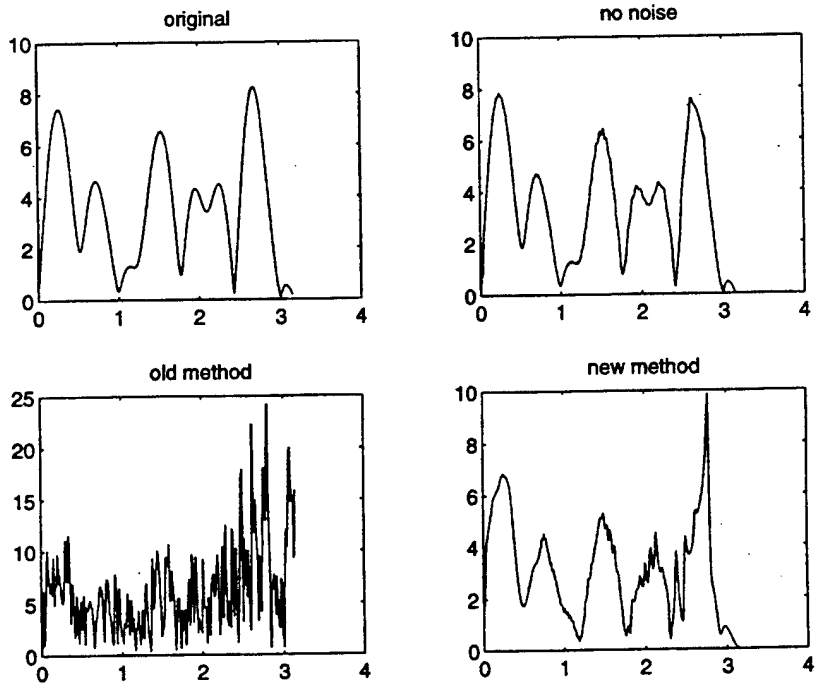

Fig. 13. System identification examples: Original spectrum $|H(\omega)|$; identified spectrum without additive noise using the chirp signal; conventional method with additive noise of SNR $= -0.4$ dB; new method with additive noise of SNR $= -0.4$ dB.

using pseudo-random sequences

$$H_{\text{old}_2}(\omega) = \frac{S_{y\tilde{x}}(\omega)}{S_{\tilde{x}\tilde{x}}(\omega)} \tag{4.3}$$

where $\tilde{x}[n]$ is a pseudo-random sequence.

Fig. 12 shows their performances, where 200 tests are used for the mean SNR curves for the system spectrum versus the original SNR. Our new method performs much better than others. Surprisingly, even for the conventional cross spectral method, the chirp signal in (3.1) outperforms pseudo-random signals by about 6 dB. In Fig. 13, some identification examples are shown, where the original SNR is -0.4 dB. As a remark, all system identification calculations used in this paper are based on the Matlab Signal Processing Toolbox.

V. Conclusion

In this paper, we proposed a system identification method. The proposed method is based on transmitting chirp signals

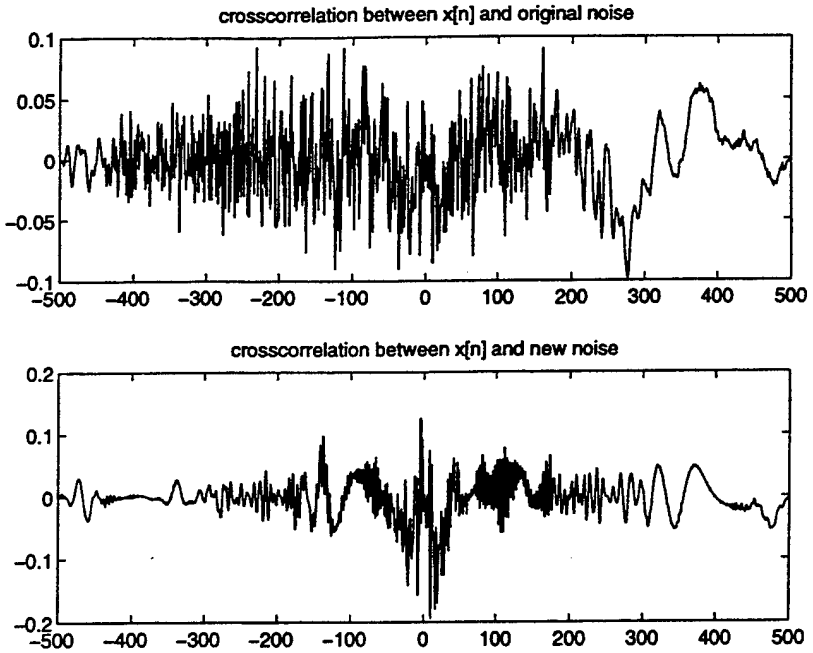

Fig. 14. Cross correlations between the new noise $\tilde{v}[n]$ (SNR $= 0.74$ dB) and the signal $x[n]$ and the original noise $v[n]$ and the signal $x[n]$ (SNR $= -6.4$ dB).

and denoising followed by the conventional identification method. The denoising method is based on time-variant filtering in the joint time-frequency (TF) domain. Since transmitted signals are chirp-type signals, they are well-localized in the TF domain, and one is usually able to see their patterns in the TF domain, even in a very low SNR environment. Due to this property, a significant SNR increase after a time-variant filtering can be achieved. Our numerical simulations were performed to illustrate this theory. The simulations done in this paper were used simply for showing the potential performance of the new approach based on time-frequency analysis and synthesis techniques in very low SNR environment. Several further improvements are possible. They are

i) to use more sophisticated detection methods in designing masks D for the iterative time-variant filter;
ii) to search the optimal reference signal $x[n]$ so that its Fourier spectrum is as far away from 0 as possible and it localizes in the TF domain as much as possible;
iii) to use more sophisticated existing system identification methods, such as the method recently proposed in [1] by Shalvi and Weinstein, where the additive noise $v[n]$ in the system model is not necessarily independent of the signal $x[n]$.

The reason for mentioning iii) here is because of the following argument. Since a joint TF domain filter that usually depends on the signal $x[n]$ is used, the new noise $\tilde{v}[n]$ after denoising and the transmitted signal $x[n]$ may have similar TF characteristics, and therefore, they may be correlated, in particular, when the original SNR is too low. Such an example is shown in Fig. 14, where the original SNR $= -6.4$ dB and the SNR $= 0.74$ dB after the second iteration of the time-variant filtering. From Fig. 14, one can clearly see that the correlation between the new noise $\tilde{v}[n]$ after denoising and the signal $x[n]$ exists, whereas it does not exist between the original noise $v[n]$ and $x[n]$. It should be observed from our numerous numerical examples that this phenomenon only happens when the original SNR is very low.

ACKNOWLEDGMENT

The author would like to thank Prof. G. Arce and Prof. C. Boncelet for their encouragement and various discussions on the subject. He also wishes to thank the referees and the associate editor for bringing [28]–[33] to his attention.

REFERENCES

[1] O. Shalvi and E. Weinstein, "System identification using nonstationary signals," *IEEE Trans. Signal Processing*, vol. 44, pp. 2055–2063, Aug. 1996.
[2] L. Ljung, *System Identification Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[3] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[4] K. Pahlavan and J. W. Matthews, "Performance of adaptive matched filter receivers over fading multipath channels," *IEEE Trans. Commun.*, vol. 38, pp. 2106–2113, Dec., 1990.
[5] S. Qian and D. Chen, *Joint Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[6] X.-G. Xia and S. Qian, "Discrete Gabor transform based time-variant filters," preprint, 1996.
[7] ——, "An iterative algorithm for time-variant filtering in the discrete Gabor transform domain," in *Proc. IEEE ICASSP'97*, Munich, Germany, Apr. 1997.
[8] C. Wilcox, "The synthesis problem for radar ambiguity functions," Tech. Summary Rep. 157, Math. Res. Cent., Univ. Wisconsin, Madison, Apr. 1960.
[9] G. F. Boudreaux-Bartels and T. W. Parks, "Time-varying filtering and signal estimation using Wigner distribution synthesis techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 442–451, June 1986.
[10] S. Farkash and S. Raz, "Time-variant filtering via the Gabor expansion," in *Signal Processing V: Theories and Applications*. New York: Elsevier, 1990, pp. 509–512.
[11] F. Hlawatsch and W. Krattenthaler, "Bilinear signal synthesis," *IEEE Trans. Signal Processing*, vol. 40, pp. 352–363, Feb. 1992.

[12] W. Kozek and F. Hlawatsch, "A comparative study of linear and nonlinear time-frequency filters," in *Proc. IEEE Int. Symp. Time-Freq. Time Scale Anal.*, Victoria, B.C., Canada, Oct. 1992, pp. 163–166.
[13] F. Hlawatsch, A. H. Costa, and W. Krattenthaler, "Time-frequency signal synthesis with time-frequency extrapolation and don't-care regions," *IEEE Trans. Signal Processing*, vol. 42, pp. 2513–2520, Sept. 1994.
[14] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55–69, Feb. 1980.
[15] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[16] M. J. Bastiaans, "Gabor's expansion of a signal into Gaussian elementary signals," *Proc. IEEE*, vol. 68, pp. 594–598, Apr. 1980.
[17] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Process.*, vol. 21, pp. 207–220, 1990.
[18] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Trans. Signal Processing*, vol. 41, pp. 2429–2438, July 1993.
[19] ———, "Optimal biorthogonal analysis window function for discrete Gabor transform," *IEEE Trans. Signal Processing*, vol. 42, pp. 694–697, Mar. 1994.
[20] S. Qian, K. Chen, and S. Li, "Optimal biorthogonal functions for finite discrete-time Gabor expansion," *Signal Process.*, vol. 27, pp. 177–185, 1992.
[21] A. J. E. M. Janssen, "Duality and biorthogonality for discrete-time Weyl-Heisenberg frames," RWR-518-RE-94001-ak unclassified rep. 002/94, Philips Res. Lab., Eindhoven, The Netherlands, 1994.
[22] X.-G. Xia, "On characterization of the optimal biorthogonal window functions for Gabor transforms," *IEEE Trans. Signal Processing*, vol. 44, pp. 133–136, Feb. 1996.
[23] L. Cohen, "Time-frequency distributions—A review," *Proc. IEEE*, vol. 77, pp. 941–981, July 1989.
[24] ———, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[25] F. Hlawatsch and G. F. Bourdeaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Mag.*, vol. 9, pp. 21–67, Apr. 1992.
[26] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, vol. 8, pp. 14–38, Oct. 1991.
[27] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, Sept. 1990.
[28] C. C. Goodwin and L. Payne, *Dynamic System Identification Experiment Design and Data Analysis*. London, U.K.: Academic, 1977.
[29] R. K. Mehra, "Optimal input signals for parameter estimation in dynamic systems-survey and new results," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 753–768, 1974.
[30] E. Rafajlowic, "Unbounded power input signals in optimum experiment design for parameter estimation in linear system," *Int. J. Contr.*, vol. 40, pp. 383-391, 1984.
[31] B. Kuszta and N. K. Sinha, *Modeling and Identification of Dynamic Systems*. New York: Van Nostrand Reinhold, 1983.
[32] M. B. Zarrop, *Optimal Experiment Design for Dynamic System Identification*. Berlin-New York: Spinger-Verlag, 1979.
[33] S. Palavajjhala, R. L. Motard, and B. Joseph, "Process identification using discrete wavelet transform: design of prefilters," *AIChe J.*, vol. 42, pp. 777–790, Mar. 1996.

Xiang-Gen Xia (M'97) received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, the M.S. degree in mathematics from Nankai University, Tianjin, China, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1983, 1986, and 1992, respectively.

He was a Lecturer at Nankai University from 1986 to 1988, a Teaching Assistant at the University of Cincinnati, Cincinnati, OH, from 1988 to 1990, a Research Assistant at the University of Southern California from 1990 to 1992, and a Research Scientist at the Air Force Institute of Technology, Wright-Patterson AFB, OH, from 1993 to 1994. He was a Senior/Research Staff Member at Hughes Research Laboratories, Malibu, CA, from 1995 to 1996. In September 1996, he joined the Department of Electrical Engineering, University of Delaware, Newark, where he is currently an Assistant Professor. His current research interests include communication systems including equalization and coding, wavelet transform and multirate filterbank theory and applications, time-frequency analysis and synthesis and numerical analysis and inverse problems in signal/image processing.

Dr. Xia received the National Science Foundation Faculty Early Career Development (CAREER) Program Award in 1997. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is also a member of the American Mathematical Society.

A Family of Pulse-Shaping Filters with ISI-Free Matched and Unmatched Filter Properties

Xiang-Gen Xia

Abstract—**The raised-cosine pulse-shaping filter plays an important role in digital communications due to its intersymbol interference (ISI)-free property. The ISI-free property holds after matched filtering is performed. In this letter, we propose a new family of pulse-shaping filters. These filters are ISI free with or without matched filtering. Using these new pulse-shaping filters, the computational load, and therefore the hardware cost in demodulation for modem design, might be reduced in some applications.**

Index Terms—**ISI-free property, matched and unmatched filtering, pulse-shaping filters.**

I. Introduction

THE raised-cosine filter

$$H(\omega) = \begin{cases} 1, & 0 \le \omega \le \frac{\pi}{T_s}(1-\alpha) \\ \cos^2\left[\frac{T_s}{4\alpha}\left(\omega - \frac{\pi(1-\alpha)}{T_s}\right)\right], & \frac{\pi}{T_s}(1-\alpha) \le \omega \le \frac{\pi}{T_s}(1+\alpha) \\ 0, & \omega > \frac{\pi}{T_s}(1+\alpha) \end{cases} \quad (1)$$

plays an important role in digital communication systems. It has been used extensively in modem design for both wireline and radio systems. This is mainly due to its intersymbol interference (ISI)-free property, i.e.,

$$h(nT_s) = \delta(n) = \begin{cases} 1, & n = 0 \\ 0, & n = \pm 1, \pm 2, \cdots \end{cases}$$

where $H(\omega)$ and $h(t)$ are the frequency and the time response functions, respectively. There have been extensive discussions of this topic; see, for example, [1]–[4].

Since the ISI-free property holds after the matched filtering is performed for the received signal, the frequency response $G(\omega)$ of the transmitted waveform $g(t)$ should be the square root of $H(\omega)$ in (1), i.e.,

$$G(\omega) = \sqrt{H(\omega)} \quad \text{and} \quad g(t) = \mathcal{F}^{-1}(G(\omega)) \quad (2)$$

where $\mathcal{F}$ stands for the Fourier transform and $\mathcal{F}^{-1}$ means its inverse. The matched filtering plays two roles here. One is low-pass filtering that reduces the noise, and the other is ISI reduction due to the ISI-free property of the raised-cosine filters. Since the length of these filters is not short, the hardware implementation cost in current modem systems is significant. However, it may occur in practice that, for some users, the matched filtering is used purely for reducing the ISI. In this case, if the transmitted signal is already ISI free, the matched filtering may not be necessary. The question then becomes whether it is possible to construct pulse shaping filters $G(\omega)$ at the transmitter so that both the transmitted signal and the signal after matched filtering are ISI free, i.e.,

$$g(nT_s) = \delta(n) \quad \text{and} \quad h(nT_s) = \delta(n)$$

where $h(t)$ is the time-domain waveform of $H(\omega) = |G(\omega)|^2$.

In this letter, we will positively answer this question by proposing a family of such pulse-shaping filters.

II. A New Family of Pulse-Shaping Filters

In this section, we present a new family of real-valued pulse-shaping filters which have ISI-free properties with or without matched filtering.

Let $g(t)$ denote the waveform in the time domain to be transmitted, and let $G(\omega)$ denote its Fourier transform. Let $h(t)$ be the waveform in the time domain after the matched filtering of $g(t)$ is performed, and let $H(\omega)$ denote its Fourier transform. Then, $H(\omega) = |G(\omega)|^2$. Without loss of generality, from now on, we assume $T_s = 1$. The ISI-free property for the waveform $g(t)$ is

$$g(n) = \delta(n), \qquad n \in \boldsymbol{Z}$$

where $\boldsymbol{Z}$ is the set of all integers. This is equivalent to

$$\sum_n G(\omega + 2n\pi) = 1. \quad (3)$$

The ISI-free property for the waveform $h(t)$ is

$$h(n) = \delta(n), \qquad n \in \boldsymbol{Z}$$

which is equivalent to

$$\sum_n |G(\omega + 2n\pi)|^2 = 1. \quad (4)$$

We want to construct real-valued $g(t)$ that satisfies (3) and (4).

Let $\nu(x)$ be a continuous function such that

$$\nu(x) = \begin{cases} 0, & x \le 0 \\ 1, & x \ge 1 \end{cases} \quad (5)$$

and

$$\nu(x) + \nu(1-x) = 1, \qquad x \in \boldsymbol{R} \quad (6)$$

where $\boldsymbol{R}$ is the set of all real numbers. An example of such a $\nu(x)$ is

$$\nu(x) = \begin{cases} 0, & x \le 0 \\ x^4(35 - 84x + 70x^2 - 20x^3), & 0 \le x \le 1 \\ 1, & x \ge 1 \end{cases} \quad (7)$$

Paper approved by K. Townsend, the Editor for Computer-Aided Design of Communications Systems of the IEEE Communications Society. Manuscript received January 21, 1997; revised May 5, 1997. This work was supported in part by an initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

The author is with the Department of Electrical Engineering, University of Delaware, Newark, DE 19716-3130 USA.

Publisher Item Identifier S 0090-6778(97)07267-X.

which has almost fourth-order smoothness. The simplest form for such ν is

$$\nu(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 \le x \le 1 \\ 1, & x \ge 1 \end{cases}$$

which is only continuous, but not differentiable.

We determine $g(t)$ by constructing its Fourier transform $G(\omega)$:

$$G(\omega) = \begin{cases} 1, & |\omega| \le \frac{2}{3}\pi \\ \frac{1}{2}\left(1 + e^{j\pi\nu((3/2\pi)\omega - 1)}\right), & \frac{2}{3}\pi < \omega < \frac{4}{3}\pi, \\ \frac{1}{2}\left(1 - e^{j\pi\nu((3/2\pi)(\omega+2\pi)-1)}\right), & -\frac{4}{3}\pi < \omega < -\frac{2}{3}\pi \\ 0, & |\omega| \ge \frac{4}{3}\pi. \end{cases} \tag{8}$$

Notice that the parameter function ν controls the width of the transfer band of the filter $G(\omega)$. The smoothness of the function ν determines the speed of the waveform decay of $g(t)$ in the time domain, i.e., the length of the filter. The smoother ν is implies the shorter the filter $g(t)$ will be.

Theorem 1: The pulse-shaping filters $g(t)$ defined by (8) satisfy the following properties.

1) They are real valued.
2) They are ISI free by themselves, i.e., $g(n) = \delta(n)$.
3) They are ISI free after matched filtering is performed, i.e., $h(n) = \delta(n)$.

Proof: To prove 1), we only need to prove $G^*(-\omega) = G(\omega)$ for $2\pi/3 < |\omega| < 4\pi/3$

$$\begin{aligned} G^*(-\omega) &= \tfrac{1}{2}\left(1 - e^{-i\pi\nu((3/2\pi)(-\omega+2\pi)-1)}\right) \\ &= \tfrac{1}{2}\left(1 - e^{-i\pi\nu(2-(3/2\pi)\omega)}\right) \\ &\overset{1}{=} \tfrac{1}{2}\left(1 - e^{-i\pi(1-\nu(-1+(3/2\pi)\omega))}\right) \\ &= \tfrac{1}{2}\left(1 + e^{i\pi\nu(3/2\pi)\omega - 1)}\right) \\ &= G(\omega) \end{aligned}$$

where step 1 is from (6).

To prove 2), we only need to prove (3). The form of $G(\omega)$ in (8) satisfies (3) for $2\pi/3 < \omega < 4\pi/3$

$$\sum_n G(\omega + 2n\pi) = G(\omega) + G(\omega - 2\pi) = 1.$$

This proves 2).

The property 3) can be similarly proved. □

The frequency responses $H(\omega)$ and $G(\omega)$ for the above new pulse-shaping filters in (8) with the ν function in (7), and the raised-cosine filter with $\alpha = 1/3$ in (1) and its square root are illustrated in Fig. 1.

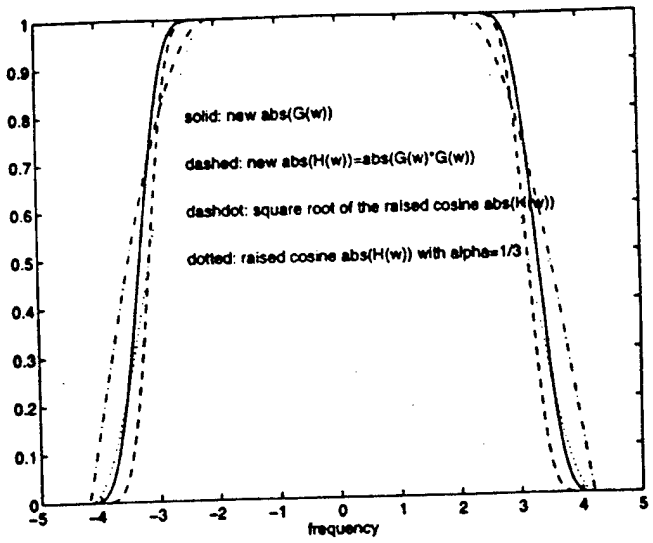

Fig. 1. The frequency responses $|H(\omega)|$ and $|G(\omega)|$ for the new pulse shaping and the raised cosine filters with $\alpha = 1/3$.

III. Conclusion

In this letter, we proposed a new family of pulse-shaping filters. These pulse-shaping filters are ISI free with or without matched filtering at the receiver. This property may reduce the hardware cost in designing modem systems in some applications where the low-pass (bandpass) filtering is performed before the matched filtering. It should be noticed that, although the new pulse-shaping filters are real valued, they are not linear phase.

Acknowledgment

The author would like to thank Dr. C.-M. Lo at Texas Instruments Incorporated for various discussions on modem design issues. He also wishes to thank the referees for their detailed comments that have improved the clarity of this paper.

References

[1] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 1994.
[2] K. Feher, *Wireless Digital Communications*. Upper Saddle River, NJ: Prentice-Hall, 1995.
[3] S. J. Maeng and B. G. Lee, "A design of linear-phased IIR Nyquist filters," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 167–175, Jan. 1995.
[4] P. P. Vaidyanathan and T. Q. Nguyen, "Eigenfilters: A new approach to least-squares FIR filter design and applications including Nyquist filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 11–23, Jan. 1987.

Inverse Problems 13 (1997) 1641–1661. Printed in the UK PII: S0266-5611(97)83932-9

A method with error estimates for band-limited signal extrapolation from inaccurate data

Xiang-Gen Xia§ and M Zuhair Nashed‖

† Department of Electrical Engineering, University of Delaware, Newark, DE 19716, USA
‡ Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA

Received 6 May 1997

Abstract. In this paper, we consider the problem of extrapolation of a band-limited signal outside a fixed interval from its (approximate or contaminated) values in that interval. We propose a new extrapolation method that estimates the error between the extrapolated and true values, and which also resolves the ill-posedness of the problem. The method is called a modified minimum norm solution (MMNS) method. Both the continuous MMNS and its discretization are studied. The error estimates hold for some classes of band-limited signals, when the maximum magnitude of the data error is known. These classes of band-limited signals are also characterized.

1. Introduction

Let f be a finite energy signal, i.e. $f \in L^2(\mathbb{R})$. Its Fourier transform $\hat{f}$ is defined by

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{it\omega}\,dt. \tag{1.1}$$

If there exists a positive number Ω such that $\hat{f}(\omega) = 0$ when $|\omega| > \Omega$, f is called Ω *band limited*. An Ω band-limited signal f can be represented by its inverse Fourier transform:

$$f(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \hat{f}(\omega)e^{-it\omega}\,d\omega. \tag{1.2}$$

It is known (see for example [1]) that a band-limited signal f is the restriction to the real line $\mathbb{R}$ of an entire function defined on the complex plane $\mathbb{C}$. Therefore, in theory, f is determined everywhere by its values on an interval no matter how small this interval is. This motivates the following band-limited signal extrapolation problem.

How does one practically extrapolate an Ω band-limited signal f outside an interval $[-T, T]$ when $f(t)$ is given for $t \in [-T, T]$ with a certain contamination error?

The above extrapolation problem is interesting not only in theory but also in many applications, such as spectral estimation (Papoulis [25]) and limited-angle tomography in medical image reconstruction (Natterer [24]), where only limited observation data are available.

Since f is analytic, a trivial solution for the problem is to compute the derivatives $f^{(n)}$ at $t = 0$ by using the values of f in $[-T, T]$ and then use the Taylor expansion. However,

§ E-mail address: xxia@ee.udel.edu
‖ E-mail address: nashed@math.udel.edu

0266-5611/97/061641+21$19.50

this method is extremely unstable due to the instability of the derivative computations. Numerical differentiation is an ill-posed problem and the degree of ill-posedness (which can be made precise using Sobolev negative norms) increases with the order of differentiation. Therefore, researchers have been seeking other methods. Since the early 1970s there has been considerable interest in this area, for example [4–8, 11–17, 24–30, 32–36, 38-40]. Since the problem itself is basically an inverse problem, it has been recognized that the existing extrapolation methods are generally unstable in terms of inaccurate data. The extrapolated values can change dramatically when the given data in an interval change slightly, see for example [27]. There are also many modified algorithms that have been proposed to improve the extrapolation performance. However, to the best of our knowledge there is no extrapolation algorithm with which one is able to estimate the error between the extrapolated and true values outside the given interval $[-T, T]$ for any nontrivial class of Ω band-limited signals, when the given data are inaccurate.

In this paper, we propose a new extrapolation method for band-limited signals that we call a *modified minimum norm solution* (MMNS) method. With the MMNS method we are able to estimate the error between the extrapolated and true values for some nontrivial classes of band-limited signals, when the maximum magnitude of the error of the given inaccurate data in a certain interval is known. This paper is organized as follows. In section 2 we study the MMNS method for continuous-time signals. In section 3 we study the MMNS method for discrete-time signals, which is a discretization of the method in section 2. In section 4 we present tractable characterizations of the classes of band-limited signals studied in sections 2 and 3. In section 5 we make several remarks.

2. Band-limited signal extrapolation in the continuous-time domain

In this section, we study the MMNS method for continuous-time band-limited signals. Without loss of generality, in what follows we assume $\Omega = 2\pi$ and $T = 1$ although we continue to use Ω and T to emphasize where they appear. We also assume $f_\epsilon = f + \eta$ where η is the error signal that is continuous in time and $|\eta(t)| \leqslant \epsilon$ for $t \in [-T, T]$, and $f_\epsilon(t)$ for $t \in [-T, T]$ are the given data. By normalization, we may assume that the maximal error magnitude $\epsilon < 1$.

We first introduce some notation. Let $L^2[-D, D]$ denote the space of all signals f that satisfy

$$\|f\|_{(D)} \triangleq \left(\int_{-D}^{D} |f(t)|^2 \, dt \right)^{1/2} < \infty$$

where D is a positive number or ∞.

Let $\mathcal{BL}$ denote all Ω band-limited signals. For $\gamma \geqslant 0$, let $\mathcal{BL}_\gamma$ denote all Ω band-limited signals $f \in \mathcal{BL}$ that satisfy the following condition.

For any $\delta > 0$, there exists a signal $g_\delta \in L^2[-T, T]$ such that

$$\hat{f}_\delta(\omega) \triangleq \frac{1}{2\pi} \int_{-T}^{T} g_\delta(t) e^{it\omega} \, dt \tag{2.1}$$

satisfies the following two properties:

$$\|\hat{f} - \hat{f}_\delta\|_{(\Omega)} \leqslant \delta \tag{2.2}$$

and

$$\|\hat{f}_\delta\|_{(\infty)} \leqslant C\delta^{-\gamma} \tag{2.3}$$

where C is a constant that is independent of δ and γ, and $\hat{f}$ is the Fourier transform of f.

The physical meaning of the above subspace of all Ω band-limited signals is as follows. For an Ω band-limited signal f, its Fourier transform $\hat{f}$ is supported in $[-\Omega, \Omega]$ and $\hat{f} \in L^2[-\Omega, \Omega]$. The correspondence between the space $\mathcal{BL}$ of all Ω band-limited signals and the space $L^2[-\Omega, \Omega]$ of all finite L^2 norm signals defined on $[-\Omega, \Omega]$ is one-to-one and onto. Therefore, for a general Ω band-limited signal f its Fourier transform $\hat{f}$ may not have any smoothness property. The subspace $\mathcal{BL}_\gamma$ contains all Ω band-limited signals f with the following properties.

(i) The Fourier transform $\hat{f}$ can be approximated in the L^2 sense by a family $\{\hat{f}_\delta\}$ of T band-limited signals (entire functions of exponential order). This approximation holds inside the frequency band of f, i.e. the support $[-\Omega, \Omega]$ of $\hat{f}$.

(ii) The L^2 norms on the whole real line of the signals in the family $\{\hat{f}_\delta\}$ may not be uniformly bounded, but the rate of the *divergence* is not arbitrary. Rather the rate is related to the rate of the *convergence* of $\{\hat{f}_\delta\}$ in $L^2[-\Omega, \Omega]$ to $\hat{f}$ as $\delta \to 0$.

In this approximations framework, what is gained is the smoothness while what is lost is the boundedness of the family of L^2 norms on the real line. This trade-off is similar to the bandwidth and the timewidth trade-off [29, 30]. More precise interpretation and characterization of the above subspace will be given in section 4.

For the maximal error magnitude ϵ mentioned at the beginning of this section and any number $\lambda \geqslant 0$, let $\mathcal{BT}_{\epsilon,\lambda}$ denote the set of all signals $g \in L^2[-T, T]$ such that

$$\left| \frac{1}{2\pi^2} \int_{-T}^{T} \frac{\sin 2\pi(s-t)}{s-t} g(s)\,ds - f_\epsilon(t) \right| \leqslant \lambda \qquad \text{for } t \in [-T, T]. \tag{2.4}$$

The basic idea for this subspace is to find signals in a neighbourhood of the inaccurate data signal $f_\epsilon(t)$ for $t \in [-T, T]$ such that the Fourier transforms of these signals are T band limited.

For $\lambda > \epsilon$, let $g_{\epsilon,\lambda}$ be the unique element (the existence and uniqueness will be shown in lemma 2) in $\mathcal{BT}_{\epsilon,\lambda}$ that has the minimum norm:

$$\|g_{\epsilon,\lambda}\|_{(T)} = \min\{\|g\|_{(T)};\, g \in \mathcal{BT}_{\epsilon,\lambda}\}. \tag{2.5}$$

Let

$$f_{\epsilon,\lambda}(t) = \frac{1}{2\pi^2} \int_{-T}^{T} \frac{\sin 2\pi(s-t)}{s-t} g_{\epsilon,\lambda}(s)\,ds \tag{2.6}$$

which is called the MMNS of the continuous-time band-limited signal extrapolation problem. We now have the following error analysis for the above MMNS.

Theorem 1. *Let $f_{\epsilon,2\epsilon}$ be defined by (2.6) with the constant $\lambda = 2\epsilon$. If $f \in \mathcal{BL}_\gamma$ for some number γ with $0 \leqslant \gamma < \frac{1}{2}$, then*

$$|f_{\epsilon,2\epsilon}(t) - f(t)| \leqslant C\epsilon^{(1-2\gamma)/3} \qquad \textit{for all } t \in \mathbb{R} \tag{2.7}$$

where C is a constant independent of ϵ and γ.

Before we prove theorem 1, we establish two lemmas. We first recall the following known results from operator theory of ill-posed problems. Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be two Hilbert spaces, and K be a bounded linear operator from $\mathbb{H}_1$ to $\mathbb{H}_2$. Let K^* denote the adjoint of the operator K and $K^\dagger$ be the generalized inverse of K (see [9, 19, 20]). Let $\mathcal{R}(K^*)$ denote the range of the operator K^*.

We recall that the (Moore–Penrose) generalized inverse $K^\dagger$ of the operator K is characterized by the following extremal property. For any g in the domain $\mathcal{D}(K^\dagger) = \mathcal{R}(K) + \mathcal{R}(K)^\perp$, the element $K^\dagger g$ is the minimal norm least-squares solution of the operator

equation $Kf = g$. If $\mathcal{R}(K)$ is nonclosed, which is the case, for example, when K is a compact operator with infinite-dimensional range, then the operator $K^\dagger$ is unbounded, so the problem is ill-posed. The well known Tikhonov regularization uses the approximation

$$x_\alpha = (K^*K + \alpha I)^{-1}K^*g \qquad \alpha > 0$$

where I is the identity operator. It is well known that

$$\lim_{\alpha \to 0} x_\alpha = K^\dagger g \qquad \text{for } g \in \mathcal{D}(K^\dagger).$$

Without any 'smoothness' assumption on $K^\dagger g$, it is not possible in general to estimate the rate of convergence of x_α to $K^\dagger g$ or to obtain an error estimate $\|x_\alpha - K^\dagger g\|$ for fixed $\alpha > 0$. In what follows we will use the following proposition (see, e.g., [10, 18]) which states that if $K^\dagger g \in \mathcal{R}(K^*)$, a kind of smoothness condition, then an error estimate holds.

Proposition 1. *If $K^\dagger g \in \mathcal{R}(K^*)$, say $K^\dagger g = K^*g^*$ for some $g^* \in \mathbb{H}_2$, then*

$$\|K^\dagger g - x_\alpha\| \leqslant \sqrt{\alpha}\|g^*\|.$$

Let us consider the operator F^{-1} from $L^2[-\Omega, \Omega]$ to $L^2[-T, T]$, a restriction of the inverse Fourier transform (1.2), defined by:

$$(F^{-1}\hat{f})(t) = f(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \hat{f}(\omega)e^{-it\omega}d\omega \qquad t \in [-T, T]. \tag{2.8}$$

Then its adjoint $(F^{-1})^*$ is

$$[(F^{-1})^*g](\omega) = \frac{1}{2\pi}\int_{-T}^{T} g(s)e^{is\omega}\,ds \qquad \omega \in [-\Omega, \Omega].$$

From (2.8), $(F^{-1}\hat{f})(t) = 0$ for almost all $t \in [-T, T]$ if and only if $\hat{f}(\omega) = 0$ for almost all $\omega \in [-\Omega, \Omega]$. This implies that the null space $\mathcal{N}(F^{-1})$ of the operator F^{-1} is the zero element. This also implies that the space $\mathcal{R}((F^{-1})^*)$ is dense in $L^2[-\Omega, \Omega]$ since Closure$(\mathcal{R}((F^{-1})^*)) = \mathcal{N}(F^{-1})^\perp = L^2[-\Omega, \Omega]$. Thus we have proved the following lemma.

Lemma 1. *For any $\delta > 0$, there exists $g_\delta \in L^2[-T, T]$ such that*

$$\|\hat{f} - \hat{f}_\delta\|_{(\Omega)} \leqslant \delta$$

where

$$\hat{f}_\delta(\omega) = \frac{1}{2\pi}\int_{-T}^{T} g_\delta(s)e^{is\omega}\,ds$$

and $\hat{f}$ is the Fourier transform of f.

By lemma 1 and its implication in the time domain, it is clear that the set $\mathcal{BT}_{\epsilon,\lambda}$ defined by (2.4) is not empty when $\lambda > \epsilon$. Since the set $\mathcal{BT}_{\epsilon,\lambda}$ is closed and convex, we have proved the following.

Lemma 2. *For $\lambda > \epsilon$, there is a unique element $g_{\epsilon,\lambda}$ in $\mathcal{BT}_{\epsilon,\lambda}$ such that*

$$\|g_{\epsilon,\lambda}\|_{(T)} = \min\{\|g\|_{(T)} : g \in \mathcal{BT}_{\epsilon,\lambda}\}.$$

With the function $g_{\epsilon,\lambda}$ as in lemma 2, define

$$\bar{g}_{\epsilon,\lambda}(\omega) = \frac{1}{2\pi}\int_{-T}^{T} g_{\epsilon,\lambda}(s)e^{is\omega}\,ds. \tag{2.9}$$

Then the MMNS $f_{\epsilon,\lambda}$ in (2.6) can also be represented as

$$f_{\epsilon,\lambda}(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \bar{g}_{\epsilon,\lambda}(\omega)e^{-is\omega}d\omega.$$

With the signal $\hat{f}_\delta$ in (2.1), define

$$\bar{f}_\delta(t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \hat{f}_\delta(s)e^{-ist}ds. \tag{2.10}$$

We are now ready to prove theorem 1.

Proof of theorem 1. When $f \in \mathcal{BL}_\gamma$ for $\gamma \geqslant 0$, by (2.1), (2.2) the signal g_δ with $\delta = (2\pi/\sqrt{2\Omega})\epsilon$ satisfies

$$\|\hat{f} - \hat{f}_\delta\|_{(\Omega)} \leqslant \frac{2\pi}{\sqrt{2\Omega}}\epsilon$$

where $\hat{f}_\delta$ is related to g_δ via (2.1). In the time domain, by using the Cauchy–Schwarz inequality and the above inequality we have

$$|f(t) - f_\delta(t)| \leqslant \frac{1}{2\pi}\left|\int_{-\Omega}^{\Omega}(\hat{f}(\omega) - \hat{f}_\delta(\omega))e^{-it\omega}\,d\omega\right| \leqslant \epsilon$$

where

$$\begin{aligned} f_\delta(t) &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega}\hat{f}_\delta(\omega)e^{-it\omega}d\omega \\ &\overset{(2.1)}{=} \frac{1}{2\pi}\int_{-\Omega}^{\Omega}\frac{1}{2\pi}\int_{-T}^{T} g_\delta(s)e^{i\omega(s-t)}ds\,d\omega \\ &= \frac{1}{2\pi^2}\int_{-T}^{T}\frac{\sin 2\pi(s-t)}{s-t}g_\delta(s)\,ds \end{aligned}$$

where the convention $\Omega = 2\pi$ made at the beginning of this section is used. By the assumption

$$|f_\epsilon(t) - f(t)| \leqslant \epsilon$$

we have

$$|f_\delta(t) - f_\epsilon(t)| \leqslant 2\epsilon.$$

According to (2.4), we have proved that g_δ is in $\mathcal{BT}_{\epsilon,2\epsilon}$. Hence, by lemma 2 we obtain

$$\|g_{\epsilon,2\epsilon}\|_{(T)} \leqslant \|g_{(2\pi/\sqrt{2\Omega})\epsilon}\|_{(T)}.$$

Moreover, by (2.1) and (2.3), we have

$$\|g_{\epsilon,2\epsilon}\|_{(T)} \leqslant \|g_{(2\pi/\sqrt{2\Omega})\epsilon}\|_{(T)} \leqslant 2\pi C(2\pi/\sqrt{2\Omega})^{-\gamma}\epsilon^{-\gamma}.$$

Since

$$|f_{\epsilon,2\epsilon}(t) - f_\epsilon(t)| \leqslant 2\epsilon \qquad t \in [-T, T]$$

we have

$$|f_{\epsilon,2\epsilon}(t) - f(t)| \leqslant 3\epsilon \qquad t \in [-T, T].$$

For the signal $\bar{f}_\delta$ in (2.10) and considering (2.2) in the time domain, we have

$$|\bar{f}_\delta(t) - f(t)| \leqslant \frac{\sqrt{2\Omega}}{2\pi}\delta \qquad \text{for } t \in \mathbb{R}.$$

Therefore,

$$|f_{\epsilon,2\epsilon}(t) - \bar{f}_\delta(t)| \leqslant 3\epsilon + \frac{\sqrt{2\Omega}}{2\pi}\delta \qquad \text{for } t \in [-T, T]. \tag{2.11}$$

For $\alpha > 0$, let

$$x_\alpha = ((F^{-1})^* F^{-1} + \alpha I)^{-1}(F^{-1})^*(f_{\epsilon,2\epsilon}(t) - \bar{f}_\delta(t)).$$

By using proposition 1 with $K = F^{-1}$ and $\delta = \epsilon$, and (2.1), (2.2), we have

$$\begin{aligned}\|\bar{g}_{\epsilon,2\epsilon} - \hat{f}_\delta - x_\alpha\|_{(\Omega)} &= \|K^\dagger(f_{\epsilon,2\epsilon} - \bar{f}_\delta) - x_\alpha\|_{(\Omega)} \\ &\leqslant \sqrt{\alpha}(\|g_{\epsilon,2\epsilon}\|_{(T)} + \|g_\delta\|_{(T)}) \\ &\leqslant 2\pi C\epsilon^{-\gamma}\sqrt{\alpha},\end{aligned}$$

where C is a constant, and $\bar{g}_{\epsilon,2\epsilon} - \hat{f}_\delta = K^*(g_{\epsilon,2\epsilon} - g_\delta)$ from (2.1) and (2.9). On the other hand,

$$\|x_\alpha\|_{(\Omega)} \leqslant \frac{1}{\alpha}\frac{T\sqrt{2\Omega}}{\pi}\left(3\epsilon + \frac{\sqrt{2\Omega}}{2\pi}\delta\right) = \frac{T\sqrt{2\Omega}}{\pi}\left(3 + \frac{\sqrt{2\Omega}}{2\pi}\right)\frac{\epsilon}{\alpha}.$$

Thus,

$$\|\bar{g}_{\epsilon,2\epsilon} - \hat{f}_\delta\|_{(\Omega)} \leqslant 2\pi C\epsilon^{-\gamma}\sqrt{\alpha} + \frac{T\sqrt{2\Omega}}{\pi}\left(3 + \frac{\sqrt{2\Omega}}{2\pi}\right)\frac{\epsilon}{\alpha}.$$

Using (2.2) with $\delta = \epsilon$, we have

$$\|\bar{g}_{\epsilon,2\epsilon} - \hat{f}\|_{(\Omega)} \leqslant 2\pi C\epsilon^{-\gamma}\sqrt{\alpha} + \frac{T\sqrt{2\Omega}}{\pi}\left(3 + \frac{\sqrt{2\Omega}}{2\pi}\right)\frac{\epsilon}{\alpha} + \epsilon.$$

In the time domain, using the Cauchy–Schwarz inequality, we obtain

$$|f_{\epsilon,2\epsilon}(t) - f(t)| \leqslant \frac{\sqrt{2\Omega}}{2\pi}\left[2\pi C\epsilon^{-\gamma}\sqrt{\alpha} + \frac{T\sqrt{2\Omega}}{\pi}\left(3 + \frac{\sqrt{2\Omega}}{2\pi}\right)\frac{\epsilon}{\alpha} + \epsilon\right] \qquad \text{for } t \in \mathbb{R}.$$

Therefore, estimate (2.7) in theorem 1 can be proved by taking $\alpha = \epsilon^{2(1+\gamma)/3}$ and using the assumption $\epsilon < 1$ made at the beginning of this section. □

3. Discretization of the MMNS method

Since in practice we usually process discrete-time signals, it is very important to consider the discretization of the MMNS method proposed in section 2. To do so, we need some notation.

For any number λ with $\lambda > \epsilon$ and positive integer m, let $\mathcal{M}l^2_\lambda(2m+1)$ denote the set of $(2m+1)$-dimensional vectors $\boldsymbol{a} = \{a(k)\} \in \mathbb{C}^{2m+1}$ such that

$$\left|\frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m}\frac{\sin 2\pi\left(\frac{k}{m} - \frac{n}{m}\right)}{\frac{k}{m} - \frac{n}{m}}a(k) - f_\epsilon\left(\frac{n}{m}\right)\right| \leqslant \lambda \qquad \text{for } -m \leqslant n \leqslant m. \tag{3.1}$$

For $\lambda > \epsilon$, let $z_m^\lambda = \{z_m^\lambda(k)\}$ be the unique element (the existence and the uniqueness will be shown in lemma 4) in $\mathcal{M}l_\lambda(2m+1)$ such that

$$\|z_m^\lambda\| = \min\{\|a\|;\ a = \{a(k)\} \in \mathcal{M}l_\lambda^2(2m+1)\} \tag{3.2}$$

where

$$\|a\| \triangleq \left(\sum_{k=-m}^{m} |a(k)|^2\right)^{1/2}.$$

Finally, let

$$\Psi_m^\lambda(t) = \frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m} \frac{\sin 2\pi\left(\frac{k}{m} - t\right)}{\frac{k}{m} - t} z_m^\lambda(k). \tag{3.3}$$

Notice that, for a signal $f \in \mathcal{BL}$ and any constants $\lambda > \epsilon \geqslant 0$ and any positive integer m, we can always construct the signal $f_{\epsilon,\lambda}$ in (2.6) and the signal Ψ_m^λ in (3.3) from the given data $f_\epsilon(t)$ for $t \in [-T, T]$. In other words, the MMNS $f_{\epsilon,\lambda}$ given in (2.6) and its discretization Ψ_m^λ in (3.3) can be found for any $f \in \mathcal{BL}$ using its known values on a segment.

In practice, it is usually difficult to get the MMNS $f_{\epsilon,\lambda}$ in (2.6). A practical way to compute it is to use the discretization form that is formulated by Ψ_m^λ in (3.3). We have the following convergence of the discretization Ψ_m^λ of the MMNS.

Theorem 2. *For any constant λ with $\lambda > \epsilon$, the discretization Ψ_m^λ converges to $f_{\epsilon,\lambda}$ uniformly on compact sets of $\mathbb{R}$ when $m \to \infty$.*

It is interesting to notice that the convergence result in theorem 2 does not require any additional condition for a band-limited signal f. In order to get an error estimation for the MMNS, an additional condition, i.e. $f \in \mathcal{BL}_\gamma$, in theorem 1 is needed.

To prove theorem 2, we need several lemmas.

Lemma 3. *For each fixed $\lambda_0 > \epsilon$, there exists $M > 0$ such that, when $m > M$ and $\lambda \geqslant \lambda_0$, the set $\mathcal{M}l_\lambda^2(2m+1)$ defined in (3.1) is not empty and $\|z_m^\lambda\| \leqslant C_{\lambda_0}$, where C_{λ_0} is some positive constant and independent of m and λ with $\lambda \geqslant \lambda_0$.*

Proof. By lemma 1, for $\delta = (\lambda - \epsilon)/3$, there exists $g_\delta \in L^2[-1, 1]$ such that

$$\|\hat{f} - \hat{f}_\delta\|_{(2\pi)} \leqslant (\lambda - \epsilon)/3$$

where

$$\hat{f}_\delta(\omega) = \frac{1}{2\pi}\int_{-1}^{1} g_\delta(s)e^{is\omega}\,ds.$$

Thus,

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} \hat{f}_\delta(\omega)e^{-it\omega}\,d\omega - f(t)\right| \leqslant (\lambda - \epsilon)/(3\sqrt{\pi}) \qquad \text{for all } t \in \mathbb{R}.$$

In other words,

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\frac{1}{2\pi}\int_{-1}^{1} g_\delta(s)e^{is\omega}ds\,d\omega - f(t)\right| \leqslant (\lambda - \epsilon)/3 \qquad \text{for all } t \in \mathbb{R}. \tag{3.4}$$

Since the space of continuous functions is dense in $L^2[-1, 1]$, there exists $h_\delta \in C[-1, 1]$ such that

$$\|g_\delta - h_\delta\|_{(1)} \leqslant (\lambda - \epsilon)/3.$$

Thus,

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\frac{1}{2\pi}\int_{-1}^{1}(g_\delta(s)-h_\delta(s))e^{is\omega}\,ds\,d\omega\right| \leqslant \frac{1}{2\pi}\int_{-2\pi}^{2\pi}\frac{1}{2\pi}\int_{-1}^{1}|g_\delta(s)-h_\delta(s)|\,ds\,d\omega$$
$$\leqslant \sqrt{2}(\lambda-\epsilon)/(3\pi) < (\lambda-\epsilon)/3 \qquad \text{for all } t\in\mathbb{R}.$$

By (3.4) we have

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\frac{1}{2\pi}\int_{-1}^{1}h_\delta(s)e^{is\omega}ds\,d\omega - f_\epsilon(t)\right| \leqslant (2\lambda+\epsilon)/3 \qquad \text{for all } t\in[-1,1].$$

That is,

$$\left|\frac{1}{2\pi^2}\int_{-1}^{1}\frac{\sin 2\pi(s-t)}{s-t}h_\delta(s)ds - f_\epsilon(t)\right| \leqslant (2\lambda+\epsilon)/3 \qquad \text{for all } t\in[-1,1]. \tag{3.5}$$

Since h_δ is continuous on $[-1,1]$, the following sum

$$\frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m}\frac{\sin 2\pi(\frac{k}{m}-t)}{\frac{k}{m}-t}h_\delta\left(\frac{k}{m}\right)$$

converges uniformly to

$$\frac{1}{2\pi^2}\int_{-1}^{1}\frac{\sin 2\pi(s-t)}{s-t}h_\delta(s)\,ds$$

for $t\in[-1,1]$. Therefore, for $(\lambda-\epsilon)/3$, there exists $M>0$ such that, when $m>M$, we have

$$\left|\frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m}\frac{\sin 2\pi\left(\frac{k}{m}-\frac{n}{m}\right)}{\frac{k}{m}-\frac{n}{m}}h_\delta\left(\frac{k}{m}\right)-\frac{1}{2\pi^2}\int_{-1}^{1}\frac{\sin 2\pi(s-\frac{n}{m})}{s-\frac{n}{m}}h_\delta(s)\,ds\right|$$
$$\leqslant (\lambda-\epsilon)/3 \qquad \text{for } |n|\leqslant m.$$

Combining this with (3.5), we obtain

$$\left|\frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m}\frac{\sin 2\pi\left(\frac{k}{m}-\frac{n}{m}\right)}{\frac{k}{m}-\frac{n}{m}}h_\delta\left(\frac{k}{m}\right)-f_\epsilon\left(\frac{n}{m}\right)\right| \leqslant \lambda \qquad \text{for all } |n|\leqslant m.$$

Let $a(k)=h_\delta(\frac{k}{m})$ for $|k|\leqslant m$. Then, $\{a(k)\}\in\mathcal{M}l_\lambda^2(2m+1)$. This proves that the set $\mathcal{M}l_\lambda^2(2m+1)$ is not empty when $m>M$.

Moreover, the above M can be large enough such that, when $m>M$,

$$\frac{1}{m}\sum_{k=-m}^{m}|a(k)|^2 \leqslant \int_{-1}^{1}|h_\delta(s)|^2\,ds+1$$
$$=\int_{-1}^{1}|h_{(\lambda-\epsilon)/3}(s)|^2\,ds+1$$
$$\leqslant (\|g_{(\lambda-\epsilon)/3}\|_{(1)}+(\lambda-\epsilon)/3)^2+1.$$

Let $(\lambda-\epsilon)/3<1$ and

$$C_{\lambda_0}=\left\{(\|g_{(\lambda_0-\epsilon)/3}\|_{(1)}+1)^2+1\right\}^{1/2}.$$

Then lemma 3 is proved. □

Similar to lemma 2, since the set $\mathcal{M}l^2_\lambda(2M+1)$ is closed and convex, we prove lemma 4.

Lemma 4. *For every m and λ with $\lambda > \epsilon$, there exists a unique element*

$$z^\lambda_m = \{z^\lambda_m(k)\} \in \mathcal{M}l^2_\lambda(2m+1)$$

such that

$$\|z^\lambda_m\| = \min\{\|a\| : a = \{a(k)\} \in \mathcal{M}l^2_\lambda(2m+1)\}.$$

Recall that a family of functions of a complex variable is called a *normal family* if every sequence of the family contains a subsequence which converges uniformly on compact sets. It is known that a family of functions that is uniformly bounded in any compact set is a normal family. We use this result in the proof of the following lemma.

Lemma 5. *For each λ_0 ($> \epsilon$), the family of functions $\{\Psi^\lambda_m(t)\}_{\lambda \geqslant \lambda_0, m}$ defined in (3.3) is normal when t is extended to the complex plane $\mathbb{C}$.*

Proof. The functions Ψ^λ_m in (3.3) can be rewritten as

$$\begin{aligned}\Psi^\lambda_m(t) &= \frac{1}{4\pi^2}\frac{1}{m}\int_{-2\pi}^{2\pi} e^{-it\omega}\sum_{k=-m}^{m} e^{ik\omega/m} z^\lambda_m(k)\,d\omega \\ &= \frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\left(\frac{1}{2\pi}\frac{1}{m}\sum_{k=-m}^{m} e^{ik\omega/m} z^\lambda_m(k)\right)d\omega.\end{aligned}$$

Thus,

$$\begin{aligned}|\Psi^\lambda_m(z)| &\leqslant \frac{e^{2\pi|z|}}{4\pi^2}\int_{-2\pi}^{2\pi}\left|\frac{1}{m}\sum_{k=-m}^{m} e^{-ik\omega/m} z^\lambda_m(k)\right| d\omega \\ &\leqslant \frac{e^{2\pi|z|}}{\pi}\frac{1}{m}\sum_{k=-m}^{m}|z^\lambda_m(k)| \\ &\leqslant \frac{e^{2\pi|z|}}{\pi}\left(\frac{2m+1}{m}\right)^{1/2}\|z^\lambda_m\| \\ &\overset{\text{lemma 3}}{\leqslant} \frac{1}{\pi}\left(\frac{2m+1}{m}\right)^{1/2} C_{\lambda_0} e^{2\pi|z|} \qquad \text{for } \lambda \geqslant \lambda_0 \qquad z \in \mathbb{C}.\end{aligned}$$

This proves that the family $\{\Psi^\lambda_m\}_{\lambda \geqslant \lambda_0, m}$ is normal. □

Define

$$\phi^\lambda_m(\omega) = \frac{1}{2\pi}\frac{1}{m}\sum_{k=-m}^{m} e^{ik\omega/m} z^\lambda_m(k). \tag{3.6}$$

Lemma 6. *For each λ_0 ($> \epsilon$) the family $\{\phi^\lambda_m(z)\}_{\lambda \geqslant \lambda_0, m}$ is normal and its limit functions are 1 band limited.*

Proof. The proof of normality is similar to the proof of lemma 5 by using lemma 3. By Fatou's lemma and lemma 3, it is easy to prove that all limit functions of the family $\{\phi^\lambda_m(z)\}_{\lambda \geqslant \lambda_0, m}$ are in $L^2(\mathbb{R})$ when z is restricted to the real line $\mathbb{R}$. Therefore, by the Paley–Wiener theorem (see [1]), lemma 6 is proved. □

Lemma 7. *Let $g_{\epsilon,\lambda}$ be as defined in (2.5). For a fixed ϵ, let $h(\lambda) = \|g_{\epsilon,\lambda}\|_{(1)}$. Then the function $h(\lambda)$ is continuous for $\lambda > \epsilon$.*

Proof. Let λ_0 and λ_1 be any two positive numbers such that $\lambda_0 > \lambda_1 > \epsilon$. For any $\lambda \geqslant \lambda_1$, define

$$\bar{g}_{\epsilon,\lambda}(\omega) = \frac{1}{2\pi}\int_{-1}^{1} e^{is\omega} g_{\epsilon,\lambda}(s)\,ds.$$

Then

$$|\bar{g}_{\epsilon,\lambda}(\omega)| = \left|\frac{1}{2\pi}\int_{-1}^{1} e^{is\omega} g_{\epsilon,\lambda}(s)ds\right| \leqslant e^{|\omega|}\frac{\sqrt{2}}{2\pi}\|g_{\epsilon,\lambda}\|_{(1)} \leqslant \frac{\sqrt{2}}{2\pi}e^{|\omega|}\|g_{\epsilon,\lambda_1}\|_{(1)} \qquad \text{for } \lambda \geqslant \lambda_1.$$

This implies that the family $\{\bar{g}_{\epsilon,\lambda}(\omega)\}_{\lambda\geqslant\lambda_1}$ is normal. Similar to lemma 6, its limit functions are 1 band limited. Let $\bar{h}_{\epsilon,\lambda_0}$ be one of its limit functions. Let $\lambda(n) \to \lambda_0^+$ and suppose that the sequence $\{\bar{g}_{\epsilon,\lambda(n)}\}$ converges to $\bar{h}_{\epsilon,\lambda_0}$ uniformly on compact sets of $\mathbb{C}$. Then, there exists $h_{\epsilon,\lambda_0} \in L^2[-1,1]$ such that

$$\bar{h}_{\epsilon,\lambda_0}(\omega) = \frac{1}{2\pi}\int_{-1}^{1} e^{is\omega} h_{\epsilon,\lambda_0}(s)ds.$$

By the definition of $f_{\epsilon,\lambda(n)}$ we have

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\bar{g}_{\epsilon,\lambda(n)}(\omega)d\omega - f_\epsilon(t)\right| = |f_{\epsilon,\lambda(n)}(t) - f_\epsilon(t)| \leqslant \lambda(n) \qquad \text{for } t \in [-1,1].$$

Let $n \to \infty$ in the above inequality,

$$\left|\frac{1}{2\pi}\int_{-2\pi}^{2\pi} e^{-it\omega}\frac{1}{2\pi}\int_{-1}^{1} h_{\epsilon,\lambda_0}(s)e^{is\omega}\,d\omega\,ds - f_\epsilon(t)\right| \leqslant \lambda_0 \qquad \text{for } t \in [-1,1].$$

Thus, $h_{\epsilon,\lambda_0} \in \mathcal{BT}_{\epsilon,\lambda_0}$. Therefore,

$$\|h_{\epsilon,\lambda_0}\|_{(1)} \geqslant \|g_{\epsilon,\lambda_0}\|_{(1)}. \tag{3.7}$$

On the other hand, for any $B > 0$,

$$\begin{aligned}\int_{-B}^{B} |\bar{h}_{\epsilon,\lambda_0}(\omega)|^2 d\omega &= \lim_{n\to\infty}\int_{-B}^{B} |\bar{g}_{\epsilon,\lambda(n)}(\omega)|^2 d\omega \\ &\leqslant \overline{\lim}_{n\to\infty}\int_{-\infty}^{\infty} |\bar{g}_{\epsilon,\lambda(n)}(\omega)|^2 d\omega = \overline{\lim}_{n\to\infty}\|\bar{g}_{\epsilon,\lambda(n)}\|^2_{(\infty)} \\ &= \frac{1}{2\pi}\overline{\lim}_{n\to\infty}\|g_{\epsilon,\lambda(n)}\|^2_{(1)} \leqslant \frac{1}{2\pi}\|g_{\epsilon,\lambda_0}\|^2_{(1)}.\end{aligned}$$

Therefore,

$$\|\bar{h}_{\epsilon,\lambda_0}\|_{(\infty)} \leqslant \frac{1}{\sqrt{2\pi}}\|g_{\epsilon,\lambda_0}\|_{(1)}.$$

In other words,

$$\|h_{\epsilon,\lambda_0}\|_{(1)} \leqslant \|g_{\epsilon,\lambda_0}\|_{(1)}.$$

By (3.7) and lemma 2, we have proved that $h_{\epsilon,\lambda_0} = g_{\epsilon,\lambda_0}$. Therefore, we have proved

$$\lim_{\lambda\to\lambda_0^+} h(\lambda) = h(\lambda_0). \tag{3.8}$$

Now we want to prove that

$$\lim_{\lambda\to\lambda_0^-} h(\lambda) = h(\lambda_0). \tag{3.9}$$

Let λ_1 be any positive with $0 < \lambda_1 < \lambda_0$. Let $\{\lambda(n)\}$ be any sequence of numbers with $\lambda_1 \leqslant \lambda(n) \leqslant \lambda(n+1) \leqslant \lambda_0$ that converges to λ_0. Define

$$h_n(s) = \left(1 - \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1}\right) g_{\epsilon,\lambda_1}(s) + \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1} g_{\epsilon,\lambda_0}(s) \qquad \text{for } s \in [-1, 1]. \tag{3.10}$$

Define

$$\bar{h}_n(t) = \frac{1}{2\pi} \int_{-2\pi}^{2\pi} e^{-it\omega} \left(\frac{1}{2\pi} \int_{-1}^{1} h_n(s) e^{is\omega} ds \right) d\omega.$$

Then

$$\begin{aligned} |\bar{h}_n(t) - f_\epsilon(t)| &= \left| \frac{1}{2\pi^2} \int_{-1}^{1} \frac{\sin 2\pi(s-t)}{s-t} h_n(s)\, ds - f_\epsilon(t) \right| \\ &\leqslant \left(1 - \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1}\right) \left| \frac{1}{2\pi^2} \int_{-1}^{1} \frac{\sin 2\pi(s-t)}{s-t} g_{\epsilon,\lambda_1}(s)\, ds - f_\epsilon(t) \right| \\ &\quad + \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1} \left| \frac{1}{2\pi^2} \int_{-1}^{1} \frac{\sin 2\pi(s-t)}{s-t} g_{\epsilon,\lambda_0}(s)\, ds - f_\epsilon(t) \right| \\ &\leqslant \left(1 - \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1}\right) \lambda_1 + \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1} \lambda_0 = \lambda(n). \end{aligned}$$

This implies that $h_n \in \mathcal{BT}_{\epsilon,\lambda(n)}$.

From (3.10) we have

$$\|h_n\|_{(1)} \leqslant \left(1 - \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1}\right) \|g_{\epsilon,\lambda_1}\|_{(1)} + \frac{\lambda(n) - \lambda_1}{\lambda_0 - \lambda_1} \|g_{\epsilon,\lambda_0}\|_{(1)}.$$

Letting $n \to \infty$ we obtain

$$\overline{\lim}_{n\to\infty} \|h_n\|_{(1)} \leqslant \|g_{\epsilon,\lambda_0}\|_{(1)}.$$

Since we have proved that $h_n \in \mathcal{BT}_{\epsilon,\lambda(n)}$,

$$\|g_{\epsilon,\lambda(n)}\|_{(1)} \leqslant \|h_n\|_{(1)}.$$

This proves that

$$\overline{\lim}_{n\to\infty} \|g_{\epsilon,\lambda(n)}\|_{(1)} \leqslant \|g_{\epsilon,\lambda_0}\|_{(1)}.$$

On the other hand, the following is clear:

$$\|g_{\epsilon,\lambda(n)}\|_{(1)} \geqslant \|g_{\epsilon,\lambda_0}\|.$$

Thus,

$$\lim_{n\to\infty} \|g_{\epsilon,\lambda(n)}\|_{(1)} = \|g_{\epsilon,\lambda_0}\|_{(1)}$$

that is, (3.9) is proved. This proves lemma 7. □

We are now ready to prove theorem 2.

Proof of theorem 2. By (3.3) and (3.6) we have

$$\Psi_m^\lambda(t) = \frac{1}{2\pi} \int_{-2\pi}^{2\pi} \phi_m^\lambda(\omega) e^{-it\omega} d\omega.$$

If we can prove that every limit function of the sequence $\{\Psi_m^\lambda\}$ is $f_{\epsilon,\lambda}$, theorem 2 is proved. Assume $\bar{h}_{\epsilon,\lambda}$ is a limit function of the sequence $\{\Psi_m^\lambda\}$. Without loss of generality, we may assume the sequence $\{\Psi_m^\lambda\}$ converges to $\bar{h}_{\epsilon,\lambda}$. Since the family $\{\Psi_m^\lambda\}$ for a fixed λ is normal by lemma 5, the convergence is uniform on compact sets of $\mathbb{C}$. By lemma 6, the family

$\{\phi_m^\lambda\}$ is also normal for a fixed λ. We may assume that the sequence $\{\phi_m^\lambda\}$ converges to $\hat{h}_{\epsilon,\lambda}$ uniformly on compact sets of $\mathbb{C}$ and

$$\bar{h}_{\epsilon,\lambda}(t) = \frac{1}{2\pi}\int_{-2\pi}^{2\pi} \hat{h}_{\epsilon,\lambda}(\omega)\mathrm{e}^{-\mathrm{i}t\omega}\mathrm{d}\omega.$$

By Lemma 6, there exists $\tilde{h}_{\epsilon,\lambda} \in L^2[-1, 1]$ such that

$$\hat{h}_{\epsilon,\lambda}(\omega) = \frac{1}{2\pi}\int_{-1}^{1} \mathrm{e}^{\mathrm{i}s\omega}\tilde{h}_{\epsilon,\lambda}(s)\mathrm{d}s.$$

Taking the limit as $m \to \infty$ in

$$\left|\Psi_m^\lambda\left(\frac{n}{m}\right) - f_\epsilon\left(\frac{n}{m}\right)\right| \leqslant \lambda \qquad \text{for } |n| \leqslant m$$

and using the continuity of $\bar{h}_{\epsilon,\lambda}(t)$ and $f_\epsilon(t)$ for $t \in [-1, 1]$, we obtain

$$|\bar{h}_{\epsilon,\lambda}(t) - f_\epsilon(t)| \leqslant \lambda \qquad t \in [-1, 1].$$

This proves that $\tilde{h}_{\epsilon,\lambda} \in \mathcal{BT}_{\epsilon,\lambda}$. Thus,

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)} \geqslant \|g_{\epsilon,\lambda}\|_{(1)}. \tag{3.11}$$

We next want to prove the reverse inequality.

For $\lambda > \epsilon$, choose μ such that $\lambda > \mu > \epsilon$. For this μ, we have $g_{\epsilon,\mu} \in \mathcal{BT}_{\epsilon,\lambda}$. Using the same argument as in the proof of lemma 3, for $(\lambda - \mu)/3$ there exists $\tilde{g}_{\epsilon,\mu} \in C[-1, 1]$ such that

$$\|g_{\epsilon,\mu} - \tilde{g}_{\epsilon,\mu}\|_{(1)} \leqslant \frac{\lambda - \mu}{3}.$$

Thus, if we let

$$\bar{\tilde{g}}_{\epsilon,\mu}(t) = \frac{1}{2\pi^2}\int_{-1}^{1} \frac{\sin 2\pi(s-t)}{s-t}\tilde{g}_{\epsilon,\mu}(s)\,\mathrm{d}s$$

then,

$$|\bar{\tilde{g}}_{\epsilon,\mu}(t) - f_{\epsilon,\mu}(t)| \leqslant \frac{\sqrt{2}}{\pi}\frac{\lambda - \mu}{3} \qquad \text{for } t \in [-1, 1].$$

Therefore, there exists $M > 0$ such that when $m > M$ we have

$$\left|\frac{1}{2\pi^2}\frac{1}{m}\sum_{k=-m}^{m} \frac{\sin 2\pi\left(\frac{k}{m} - \frac{n}{m}\right)}{\frac{k}{m} - \frac{n}{m}}\tilde{g}_{\epsilon,\mu}\left(\frac{k}{m}\right) - f_\epsilon\left(\frac{n}{m}\right)\right| \leqslant \frac{2\lambda + \mu}{3} < \lambda.$$

By (3.1), this implies that $\tilde{g}_{\epsilon,\mu} = \{\tilde{g}_{\epsilon,\mu}(\frac{k}{m})\} \in \mathcal{M}l_\lambda^2(2m+1)$. Therefore,

$$\|\tilde{g}_{\epsilon,\mu}\| \geqslant \|z_m^\lambda\|.$$

Thus

$$\overline{\lim}_{m\to\infty}\frac{1}{m}\|z_m^\lambda\|^2 \leqslant \overline{\lim}_{m\to\infty}\frac{1}{m}\|\tilde{g}_{\epsilon,\mu}\|^2 = \|\tilde{g}_{\epsilon,\mu}\|_{(1)}^2.$$

Therefore,

$$\left(\overline{\lim}_{m\to\infty}\frac{1}{m}\|z_m^\lambda\|^2\right)^{1/2} \leqslant \|g_{\epsilon,\mu}\|_{(1)} + \frac{\lambda - \mu}{3}. \tag{3.12}$$

On the other hand, for any $B > 0$,

$$\begin{aligned}\int_{-B}^{B} |\hat{h}_{\epsilon,\lambda}(\omega)|^2 &= \lim_{m\to\infty} \int_{-B}^{B} |\phi_m^\lambda(\omega)|^2 d\omega \\ &\leqslant \overline{\lim}_{m\to\infty} \int_{-\pi m}^{\pi m} |\phi_m^\lambda(\omega)|^2 d\omega \\ &\overset{(3.6)}{=} \overline{\lim}_{m\to\infty} \int_{-\pi m}^{\pi m} \left(\frac{1}{2\pi m} \sum_{k=-m}^{m} e^{ik\omega/m} z_m^\lambda(k)\right)\left(\frac{1}{2\pi m} \sum_{k=-m}^{m} e^{-ik\omega/m} \overline{z_m^\lambda(k)}\right) d\omega \\ &= \overline{\lim}_{m\to\infty} \int_{-\pi m}^{\pi m} \frac{1}{(2\pi)^2 m^2} |z_m^\lambda(k)|^2 d\omega \\ &= \overline{\lim}_{m\to\infty} \frac{1}{2\pi m} \sum_{k=-m}^{m} |z_m^\lambda(k)|^2.\end{aligned}$$

Therefore,

$$\|\hat{h}_{\epsilon,\lambda}\|_{(\infty)}^2 \leqslant \frac{1}{2\pi} \overline{\lim}_{m\to\infty} \frac{1}{m} \sum_{k=-m}^{m} |z_m^\lambda(k)|^2.$$

Since

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)}^2 = 2\pi \|\hat{h}_{\epsilon,\lambda}\|_{(\infty)}^2$$

we have

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)}^2 \leqslant \overline{\lim}_{m\to\infty} \frac{1}{m} \sum_{k=-m}^{m} |z_m^\lambda(k)|^2.$$

By (3.12),

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)} \leqslant \|g_{\epsilon,\mu}\|_{(1)} + \frac{\lambda - \mu}{3}.$$

Letting $\mu \to \lambda$, by the continuity of $h(\lambda)$ on (ϵ, ∞) in lemma 7, we have

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)} \leqslant \|g_{\epsilon,\lambda}\|_{(1)}.$$

By (3.11), we have proved that

$$\|\tilde{h}_{\epsilon,\lambda}\|_{(1)} = \|g_{\epsilon,\lambda}\|_{(1)}.$$

Since $\tilde{h}_{\epsilon,\lambda} \in \mathcal{BT}_{\epsilon,\lambda}$, by lemma 2, we have

$$\tilde{h}_{\epsilon,\lambda}(s) = g_{\epsilon,\lambda}(s) \qquad \text{for } s \in [-1, 1], \text{ almost surely.}$$

This proves that Ψ_m^λ converges to $f_{\epsilon,\lambda}$ as $m \to \infty$. □

4. Band-limited signal spaces $\mathcal{BL}_\gamma$

The error estimate result in theorem 1 is for band-limited signals in the spaces $\mathcal{BL}_\gamma$. The conditions in (2.1)–(2.3) defining these spaces are rather abstract. In this section, we study their properties and simplifications. To do so, let us first review the prolate spheroidal wavefunctions (see [25, 29, 30]).

Let K be the following operator

$$(Kf)(t) = \int_{-T}^{T} \frac{\sin \Omega(t-\tau)}{\pi(t-\tau)} f(\tau)\, d\tau \qquad f \in L^2[-T, T]. \tag{4.1}$$

It is clear that the operator K defined on $L^2[-T, T]$ is self-adjoint and compact. Let ϕ_k and λ_k, $k = 0, 1, 2, \ldots$, be the eigenfunctions and the corresponding eigenvalues of the operator K, respectively, such that ϕ_k, $k = 0, 1, 2, \ldots$, form an orthogonal basis for $L^2[-T, T]$ with

$$\int_{-T}^{T} \phi_j(t)\phi_k(t)\,dt = \lambda_k \delta(j-k)$$

where $\delta(n) = 1$ when $n = 0$ and $\delta(n) = 0$ otherwise. Moreover, we have

$$1 > \lambda_0 > \lambda_1 > \cdots > 0 \qquad \text{and} \qquad \lambda_k \to 0 \text{ as } k \to \infty. \tag{4.2}$$

From (4.1),

$$\phi_k(t) = \frac{1}{\lambda_k} \int_{-T}^{T} \frac{\sin \Omega(t-\tau)}{\pi(t-\tau)} \phi_k(\tau)\,d\tau \qquad t \in [-T, T] \qquad k = 0, 1, 2, \ldots. \tag{4.3}$$

Although the above eigenfunctions ϕ_k are only defined on the interval $[-T, T]$, they can be easily extended to the whole real line $\mathbb{R}$ by letting t take an arbitrary real value in formula (4.3). By doing so, it was proved in [29, 30] that the extended eigenfunctions ϕ_k for $t \in \mathbb{R}$ have the following orthonormality:

$$\int_{-\infty}^{\infty} \phi_j(t)\phi_k(t)\,dt = \delta(j-k).$$

These extended eigenfunctions ϕ_k are called the *prolate spheroidal wavefunctions* in [29, 30]. It was also proved in [29, 30] that these prolate spheroidal wavefunctions ϕ_k, $k = 0, 1, 2, \ldots$, form an orthonormal basis for the Ω band-limited signal space $\mathcal{BL}$. Thus, any $f \in \mathcal{BL}$ can be expanded as

$$f(t) = \sum_{k=0}^{\infty} a_k \phi_k(t) \qquad t \in \mathbb{R} \tag{4.4}$$

where

$$a_k = \int_{-\infty}^{\infty} f(t)\phi_k(t)\,dt = \frac{1}{\lambda_k} \int_{-T}^{T} f(t)\phi_k(t)\,dt \tag{4.5}$$

and

$$\|f\|_{(\infty)}^2 = \sum_{k=0}^{\infty} a_k^2 \tag{4.6}$$

and

$$\|f\|_{(T)}^2 = \sum_{k=0}^{\infty} a_k^2 \lambda_k. \tag{4.7}$$

We now have the following result.

Theorem 3. *Let f be an Ω band-limited function and have the expansion (4.4), (4.5). If*

$$\sum_{k=0}^{\infty} \frac{a_k^2}{\lambda_k^{1-2\gamma/3}} < \infty \qquad \textit{for some } \gamma,\ 0 \leqslant \gamma < \tfrac{1}{2}$$

then $f \in \mathcal{BL}_\gamma$.

Proof. For $A > 0$, let D_A be the truncation operator on $L^2(\mathbb{R})$: for $h \in L^2(\mathbb{R})$,

$$(D_A h)(t) = \begin{cases} h(t) & t \in [-A, A] \\ 0 & \text{otherwise.} \end{cases}$$

By (4.4) and (4.5), letting F denote the Fourier transform, we obtain

$$\begin{aligned} \hat{f}(\omega) = Ff(\omega) &= \sum_{k=0}^{\infty} a_k F\phi_k(\omega) \\ &= \sum_{k=0}^{\infty} a_k D_\Omega F D_T \phi_k(t)/\lambda_k \\ &= D_\Omega \sum_{k=0}^{\infty} \frac{a_k}{\lambda_k} F D_T \phi_k(t). \end{aligned}$$

Let

$$\hat{f}_n = \sum_{k=0}^{n} a_k F\phi_k = D_\Omega \sum_{k=0}^{n} \frac{a_k}{\lambda_k} F D_T \phi_k.$$

Then

$$f_n = \sum_{k=0}^{n} a_k \phi_k$$

and

$$\|f_n - f\|_{(\infty)}^2 = \sum_{k=n+1}^{\infty} a_k^2$$

$$\|\hat{f}_n - \hat{f}\|_{(\Omega)}^2 = 2\pi \sum_{k=n+1}^{\infty} a_k^2$$

and

$$\hat{f}_n = D_\Omega F\left(D_T \sum_{k=0}^{n} \frac{a_k}{\lambda_k} \phi_k\right).$$

Let

$$g_n = 2\pi D_T \sum_{k=0}^{n} \frac{a_k}{\lambda_k} \phi_k.$$

Then

$$\hat{f}_n = D_\Omega \frac{1}{2\pi} F g_n$$

and

$$\|g_n\|_{(T)}^2 = \sum_{k=0}^{n} \frac{a_k^2}{\lambda_k^2} \|\phi_k\|_{(T)}^2 = \sum_{k=0}^{n} \frac{a_k^2}{\lambda_k}.$$

Let

$$b_k^2 = \frac{a_k^2}{\lambda_k^{1-2\gamma/3}} \qquad k = 0, 1, 2, \ldots.$$

Then by the assumption

$$B \triangleq \sum_{k=0}^{\infty} b_k^2 < \infty$$

we have

$$\|g_n\|_{(T)}^2 = \sum_{k=0}^{n} b_k^2 \lambda_k^{-2\gamma/3}$$

and

$$\|\hat{f}_n - \hat{f}\|_{(\Omega)}^2 = 2\pi \sum_{k=n+1}^{\infty} b_k^2 \lambda_k^{1-2\gamma/3}.$$

By (4.2), for any $\delta > 0$, there exists N such that

$$\lambda_k^{1-2\gamma/3} \leqslant \delta \qquad \text{for } k \geqslant N+1$$

and

$$\lambda_k^{1-2\gamma/3} > \delta \qquad \text{for } k \leqslant N.$$

Then

$$\|\hat{f}_N - \hat{f}\|_{(\Omega)}^2 \leqslant 2\pi \sum_{k=N+1}^{\infty} b_k^2 \delta \leqslant 2\pi B\delta$$

and

$$\|g_N\|_{(T)}^2 \leqslant \sum_{k=0}^{N} b_k^2 \delta^{-\frac{2\gamma/3}{1-2\gamma/3}} \leqslant B\delta^{-\frac{2\gamma}{3-2\gamma}}.$$

For $0 \leqslant \gamma < \frac{1}{2}$, there exists a constant $C > 0$ such that

$$\delta^{\gamma} \|g_N\|_{(T)}^2 \leqslant C.$$

Let

$$\hat{f}_{\sqrt{2B\pi\delta}} = \frac{1}{2\pi} F g_N.$$

Then

$$\hat{f}_N = D_\Omega \hat{f}_{\sqrt{2B\pi\delta}}$$
$$\|\hat{f}_{\sqrt{2B\pi\delta}} - \hat{f}\|_{(\Omega)} \leqslant \sqrt{2B\pi\delta}$$

and

$$\begin{aligned}\left(\sqrt{2B\pi\delta}\right)^{\gamma} \|\hat{f}_{\sqrt{2B\pi\delta}}\|_{(\infty)} &= \left(\sqrt{2B\pi\delta}\right)^{\gamma} \frac{1}{\sqrt{2\pi}} \|g_N\|_{(T)} \\ &\leqslant B^{\gamma/2}(2\pi)^{(\gamma-1)/2}\delta^{\gamma/2}\|g_N\|_{(T)} \\ &\leqslant B^{\gamma/2}C^{1/2}(2\pi)^{(\gamma-1)/2} \qquad \text{for } 0 \leqslant \gamma < \tfrac{1}{2}.\end{aligned}$$

This proves that f satisfies (2.1)–(2.3). □

Before going to the next result, we recall a result on operator equations. Suppose that K is a compact linear operator from Hilbert space $\mathbb{H}_1$ to Hilbert space $\mathbb{H}_2$. Let $\theta_1^2 \geqslant \theta_2^2 \geqslant \cdots$ be the sequence of eigenvalues of the operator K^*K, and $v_1, v_2, \ldots$ be the associated orthonormal eigenfunction sequence. Let $\mu_n = \theta_n^{-1}$ and

$$u_n = \mu_n K v_n. \tag{4.8}$$

Then $\{u_n\}$ is an orthonormal sequence in $\mathbb{H}_2$ and

$$v_n = \mu_n K^* u_n. \tag{4.9}$$

We call the sequence $\{u_n, v_n; \mu_n\}$ *a singular system* for the operator K. Then, Picard's theorem can be stated as follows (for details, see, for example [10, 20]).

Proposition 2. *Let $K : \mathbb{H}_1 \to \mathbb{H}_2$ be a compact linear operator with singular system $\{u_n, v_n; \mu_n\}$. In order that the equation $Kz = g$ has a solution, it is necessary and sufficient that $g \in \mathrm{Ker}(K^*)^{\perp} (= \mathrm{Closure}\mathcal{R}(K))$ and*

$$\sum_{n=0}^{\infty} \mu_n^2 |\langle g, u_n \rangle|^2 < \infty$$

where $\langle , \rangle$ is the inner product on $\mathbb{H}_2$.

We now have the following result.

Theorem 4. *Assume that f is Ω band limited and with expansion (4.4), (4.5). Then:*

(i) $f \in \mathcal{BL}_\gamma$ with $\gamma = 0$ if and only if its Fourier transform $\hat{f}(\omega)$ or $-\hat{f}(-\omega)$ for $\omega \in (-\Omega, \Omega)$ is a piece of T band-limited signal;

(ii) $f \in \mathcal{BL}_\gamma$ with $\gamma = 0$ if and only if

$$\sum_{k=0}^{\infty} \frac{a_k^2}{\lambda_k} < \infty.$$

Proof of (i). 'If part': If $-\hat{f}(-\omega)$ for $\omega \in (-\Omega, \Omega)$ is a piece of T band-limited signal, then there exists $g \in L^2[-T, T]$ such that

$$\hat{f}(\omega) = \frac{1}{2\pi} \int_{-T}^{T} e^{is\omega} g(s) ds \qquad \omega \in (-\Omega, \Omega).$$

For any $\delta > 0$, let $g_\delta = g$. Then, $\hat{f}_\delta(\omega) = \hat{f}(\omega)$ for $\omega \in (-\Omega, \Omega)$. Let $C = \frac{1}{2\pi} \|g\|_{(T)}$. Then

$$\|\hat{f}_\delta - \hat{f}\|_{(\Omega)} = 0 \leqslant \delta$$

and

$$\|\hat{f}_\delta\|_{(\infty)} = \frac{1}{\sqrt{2\pi}} \|g_\delta\|_{(T)} = \frac{1}{\sqrt{2\pi}} \|g\|_{(T)} = C.$$

Thus $f \in \mathcal{BL}_\gamma$ for $\gamma = 0$.

'Only if part': If $f \in \mathcal{BL}_0$, then for every $\delta > 0$ there exists $g_\delta \in L^2[-T, T]$ such that

$$\|\hat{f}_\delta - \hat{f}\|_{(\Omega)} \leqslant \delta \qquad \text{and} \qquad \|g_\delta\|_{(T)} \leqslant 2\pi C$$

where C is a constant and

$$\hat{f}_\delta(\omega) = \frac{1}{2\pi} \int_{-T}^{T} g_\delta(s) e^{is\omega} ds.$$

Thus, the function family $\{\hat{f}_\delta\}$ is normal. In fact,

$$|\hat{f}_\delta(z)| \leqslant \frac{\sqrt{2T}}{2\pi} e^{|z|T} \|g_\delta\|_{(T)} \leqslant C\sqrt{2T} e^{T|z|} \qquad \text{for all } \delta > 0 \qquad z \in \mathbb{C}.$$

Therefore, for every sequence $\{\delta_n\}$ that tends to 0 when $n \to \infty$, there is a subsequence $\{\delta_{n_j}\}$ such that $\{\hat{f}_{\delta_{n_j}}\}$ converges to a T band-limited signal $\hat{h}$ uniformly on compact sets of $\mathbb{C}$. On the other hand,

$$\left| \frac{1}{2\pi} \int_{-\Omega}^{\Omega} e^{-it\omega} \hat{f}_{\delta_{n_j}}(\omega) d\omega - f(t) \right| \leqslant \frac{\sqrt{2\pi}}{2\pi} \delta_{n_j}.$$

Letting $j \to \infty$, we obtain

$$\frac{1}{2\pi} \int_{-\Omega}^{\Omega} e^{-it\omega} \hat{h}(\omega) d\omega = f(t).$$

This proves $\hat{h}(\omega) = \hat{f}(\omega)$ for $\omega \in (-\Omega, \Omega)$, that is f is a piece of a T band-limited signal.

Proof of (ii). Let $\mathbb{H}_1 = L^2[-T, T]$ and $\mathbb{H}_2 = \mathcal{BL}_0$. The inner product on $\mathbb{H}_2$ is the usual $L^2(\mathbb{R})$ inner product. Let K be the integral operator given in (4.1). By part (i), $K(L^2[-T, T]) = \mathcal{BL}_0$. By theorem 3, all finite linear combinations of the eigenfunctions ϕ_k are in $\mathcal{BL}_0$. Thus, Closure$(\mathcal{BL}_0) = \mathcal{BL}$ and therefore, $\mathcal{BL} =$ Closure$(\mathcal{R}(K))$, where the closure is under the usual $L^2(\mathbb{R})$ norm. Also,

$$K^* f(t) = \int_{-\infty}^{\infty} \frac{\sin \Omega(s-t)}{\pi(s-t)} f(s)\, ds \qquad \text{for } f \in \mathcal{BL}_0.$$

From (4.8) and (4.9),

$$u_n = \mu_n^2 K K^* u_n.$$

Hence, $\{\mu_n^2\}$ are eigenvalues of the operator KK^* and $\{u_n\}$ are the corresponding eigenfunctions. Since

$$K^* \phi_n(t) = \int_{-\infty}^{\infty} \frac{\sin \Omega(s-t)}{\pi(s-t)} \phi_n(s)\, ds = \phi_n(t)$$

we have

$$KK^* \phi_n = K\phi_n = \lambda_n \phi_n.$$

Thus by the completeness of the sequence $\{\phi_n\}$ we have

$$\lambda_n = \mu_n^{-2} \qquad \text{and} \qquad \phi_n = u_n.$$

By proposition 2,

$$f \in \mathcal{BL}_0 \qquad \text{iff} \sum_{n=0}^{\infty} \lambda_n^{-1} |(f, \phi_n)|^2 < \infty \qquad \text{iff} \sum_{n=0}^{\infty} \frac{a_n^2}{\lambda_n} < \infty.$$

This proves (ii). □

Combining theorems 1, 3 and 4, we have the following corollaries.

Corollary 1. *For* $0 \leqslant \gamma < \frac{1}{2}$, *if*

$$f(t) = \sum_{k=0}^{\infty} a_k \phi_k(t) \qquad \textit{and} \qquad \sum_{k=0}^{\infty} \frac{a_k^2}{\lambda_k^{1-2\gamma/3}} < \infty$$

then,

$$|f_{\epsilon,2\epsilon}(t) - f(t)| \leqslant C\epsilon^{(1-2\gamma)/3} \qquad t \in \mathbb{R}.$$

Corollary 2. *Let f be Ω band limited. If its Fourier transform $\hat{f}(\omega)$ for $\omega \in (-\Omega, \Omega)$ is a piece of a T band-limited function, then*

$$|f_{\epsilon,2\epsilon}(t) - f(t)| \leqslant C\epsilon^{1/3} \qquad t \in \mathbb{R}.$$

5. Remarks

In [14, 17], approximations of Ω band-limited signals f are considered. These authors use finite data of f on $[-T, T]$ to recover the whole f on $[-T, T]$. The optimal algorithm in the worst case for the recovery has been found in [14, 17] as follows.

Let O_m be an information operator which is a mapping $O_m : \mathcal{BL} \to \mathbb{C}^m$,

$$O_m f = (f(t_1), f(t_2), \ldots, f(t_m)).$$

An algorithm Φ is a function-valued mapping on $O_m\mathcal{BL}$. The optimal algorithm using O_m in the worst case takes the form:

$$\Phi(O_m f) = \sum_{k=1}^{m} b_k \frac{\sin \Omega(\cdot - t_k)}{\cdot - t_k}$$

where the coefficients $b_1, b_2, \ldots, b_m$ are determined by the solution of the linear system

$$\sum_{k=1}^{m} b_k \frac{\sin \Omega(t_n - t_k)}{t_n - t_k} = f(t_n) \qquad n = 1, 2, \ldots, m.$$

We can see that this is similar to the discretization of the MMNS in (3.1)–(3.3).

As we have already stated, a band-limited signal is the restriction of an entire function to the real line. But it is more than this. The Paley–Wiener theorem (see [1]) gives a direct characterization of band-limited signals; namely, a signal in $L^2(\mathbb{R})$ is 2π band limited if and only if it is the restriction of an entire function and is of exponential order on the real line. This provides a powerful property for extrapolation of band-limited signals that distinguishes the problem within the realm of analytic continuation of analytic functions, and makes finer and stable recovery results possible.

There is considerable literature on uniform and nonuniform sampling theorems for the recovery of band limited and other classes of signals from a countable set of sample values (see [2, 3, 12, 23, 37]), the simplest and most celebrated version being the Shannon–Whittaker theorem, which asserts that a π band-limited signal can be reconstructed via the cardinal series

$$f(t) = \sum_{n=-\infty}^{\infty} f(n) \frac{\sin \pi(t - n)}{\pi(t - n)}.$$

Various error estimates (truncation, jitter, amplitude, and aliasing errors) are also known. The problem of signal extrapolation from an interval (which usually has a small length) is markedly different from the reconstruction of the signal f via a sampling expansion theorem (which utilizes values of f on an appropriate infinite sequence with no accumulation point).

As we have shown $\mathcal{BL}_0$ is the range of the Hilbert–Schmidt compact linear operator (4.1) on $L^2[-T, T]$. $\mathcal{BL}_0$ is nonclosed in $L^2[-T, T]$. Nashed and Wahba [21, 22] have shown that the range of a Hilbert–Schmidt compact operator K is a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_Q$ with reproducing kernel

$$Q(t, s) = \int_{-T}^{T} K(t, u) K(s, u)\, du$$

where $K(t, u)$ is the Hilbert–Schmidt kernel. The inner product on $\mathcal{H}_Q$ is given by $\langle f_1, f_2\rangle_Q = \langle K^\dagger f_1, K^\dagger f_2\rangle$ for f_1, f_2 in $\mathcal{H}_Q$, where $K^\dagger$ is the Hilbert space (Moore–Penrose) generalized inverse. Equivalently,

$$\langle f_1, f_2\rangle_Q = \int_{-T}^{T} p_1(s)p_2(s)\,ds$$

where p_i is the element of the minimal norm which satisfies $Kp = f_i$, corresponding to f_i in $\mathcal{BL}_0$ for $i = 1, 2$. We recall that a Hilbert space $\mathbb{H}$ of functions f on an interval $\mathbb{J}$ is said to be a RKHS if all the evaluation functionals $E_t(f) = f(t)$, $f \in \mathbb{H}$, for each fixed $t \in \mathbb{J}$, are continuous. Then by the Riesz's representation theorem, for each $t \in \mathbb{J}$, there exists a unique element, call it Q_t, in $\mathbb{H}$ such that $f(t) = \langle f, Q_t\rangle$, $f \in \mathbb{H}$, where $\langle\,,\rangle$ is the inner product on $\mathbb{H}$. Let $Q(t, s) = \langle Q_s, Q_t\rangle$ for s, t in $\mathbb{J}$; this is the reproducing kernel (RK) of $\mathbb{H}$, and the space $\mathbb{H}$ with RK $Q(t, s)$ is denoted by $\mathbb{H}_Q$. The space $L^2(\mathbb{J})$ is not a RKHS.

The Paley–Wiener space $\mathcal{BL}$ of band-limited signals with band $[-\pi, \pi]$ is a RKHS with RK

$$Q(t, s) = \frac{\sin\pi(t - s)}{\pi(t - s)}.$$

In [23] it is shown that there is a strong affinity between RK Hilbert spaces and sampling theorems, and general sampling theorems were established for signals belonging to a RKHS which is also a closed subspace of the Sobolev space $\mathbb{H}^{-1}$. The preceding remarks about $\mathcal{BL}_0$ and the other related spaces being RKHS may suggest that a broader framework within which the type of extrapolation results derived in this paper may also hold.

Acknowledgments

The authors would like to thank the referees for their careful reading of this manuscript and their useful suggestions which improved the exposition.

X-GX was partially supported by an initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under grant no F49620-97-1-0253, and the National Science Foundation CAREER Program under grant MIP-9703377. MZN was partially supported by NSF grant DMS-901526.

References

[1] Boas R P 1954 *Entire Functions* (New York: Academic)

[2] Butzer P L 1983 A survey of the Whittaker–Shannon sampling theorem and some of its extensions *J. Math. Res. Exposition* **3** 185–212

[3] Butzer P L, Splettstösser W and Stens R L 1988 The sampling theorem and linear predicition in signal analysis *Jahresber. Deutsch. Math.-verein.* **90** 1–70

[4] Cadzow J C 1979 An extrapolation procedure for band-limited signals *IEEE Trans. Acoustics, Speech Signal Processing* **27** 4–12

[5] Chen D S and Allebach J P 1987 Analysis of error in reconstruction of two-dimensional signals from irregularly spaced samples *IEEE Trans. Acoustics, Speech Signal Processing* **35** 173–80

[6] Fitzgerald R M and Byrne C L 1980 Extrapolation of band-limited signals: a tutorial *Signal Processing Theory and Applications* ed M Kunt and F Coulon (Amsterdam: North-Holland) pp 175–99

[7] Gerchberg R W 1974 Super-resolution through error energy reduction *Optica Acta* **21** 709–20

[8] Golomb M and Weinberger H F 1959 Optimal approximation and error bounds *On Numerical Approximation* ed R Langer (Madison, WI: University of Wisconsin Press) pp 117–90

[9] Groetsch C W 1977 *Generalized Inverses of Linear Operators: Representations and Approximations* (New York: Dekker)

[10] Groetsch C W 1984 *The Theory of Tikhonov Regularization for Fredholm Integral Equations of the First Kind* (London: Pitman)
[11] Hurt N E 1989 *Phase Retrieval and Zero Crossings* (Boston, MA: Kluwer)
[12] Jerri A J 1977 The Shannon sampling theorem—its various extensions and applications: a tutorial review *Proc. IEEE* **65** 1565–96
[13] Kolba D P and Parks T W 1983 Optimal estimation for band-limited signals including time domain considerations *IEEE Trans. Acoustics, Speech Signal Processing* **31** 113–22
[14] Kowalski M A 1986 Optimal complexity recovery of band and energy-limited signals *J. Complexity* **2** 239–54
[15] Landau H J 1986 Extrapolating a band-limited function from its samples taken in a finite interval *IEEE Trans. Inform. Theory* **IT-32** 464–70
[16] Levi L 1966 Fitting a bandlimited signal to given points *IEEE Trans. Inform. Theory* **11** 372–6
[17] Micchelli C A and Rivlin T J 1977 A survey of optimal recovery *Optimal Estimation in Approximation Theory* ed C A Micchelli and T J Rivlin (New York: Plenum)
[18] Morozov V A 1968 The error principle in the solution of operational equations by the regularization method *USSR Comput. Math. Phys.* **8** 63–87
[19] Nashed M Z 1971 Generalized inverses, normal solvability, and iterations for singular operator equations *Nonlinear Functional Analysis and Applications* ed L B Rall (New York: Academic) pp 311–59
[20] Nashed M Z (ed) 1976 *Generalized Inverses and Applications* (New York: Academic)
[21] Nashed M Z and Wahba G 1974 Convergence rates of approximate least-squares solutions of linear and integral operator equations of the first kind *Math. Comput.* **28** 69–80
[22] Nashed M Z and Wahba G 1974 Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operator equations *SIAM J. Math. Anal.* **5** 974–87
[23] Nashed M Z and Walter G G 1991 Generalized sampling theorems for functions in reproducing kernel Hiltert spaces *Math. Control, Signal, Systems* **4** 363–90
[24] Natterer F 1986 *The Mathematics of Computerized Tomography* (Stuttgart: Teubner)
[25] Papoulis A 1975 A new algorithm in spectral analysis and band-limited extrapolation *IEEE Trans. Circuits Syst.* **22** 735–42
[26] Potter L C and Arun K S 1989 Energy concentration in band-limited extrapolation *IEEE Trans. Acoustics, Speech Signal Processing* **37** 1027–41
[27] Sanz J L C and Huang T S 1983 Some aspects of band-limited signal extrapolation: models, discrete approximations and noises *IEEE Trans. Acoustics, Speech Signal Processing* **31** 1492–501
[28] Schlebusch H J and Splettstösser W 1985 On a conjecture of J L C Sanz and T S Huang *IEEE Trans. Acoustics, Speech Signal Processing* **33** 1628–9
[29] Slepian D 1978 1978 Prolate spheroidal wave functions, Fourier analysis, and uncertainty-V: The discrete case *Bell Syst. Tech. J.* **57** 1371–430
[30] Slepian D, Pollak H O and Landau H J 1961 Prolate spheroidal wave functions I, II *Bell Syst. Tech. J.* **40** 43–84
[31] Tikhonov A V and Arsenin V Y 1977 *Solutions of Ill-posed Problems* (Washington, DC: Winston–Wiley)
[32] Xia X-G 1992 An extrapolation for general analytic signals *IEEE Trans. Signal Processing* **40** 2243–9
[33] Xia X-G, Zhang Z and Lo C M 1994 Error analysis of the MMSE estimator for multidimensional band-limited extrapolations from finite samples *Signal Processing* **36** 55–69
[34] Xia X-G, Kuo C-C J and Zhang Z 1995 Multiband signal reconstruction from finite samples *Signal Processing* **42** 273–89
[35] Xia X-G, Kuo C-C J and Zhang Z 1995 Signal extrapolation in wavelet subspaces *SIAM J. Sci. Comput.* **16** 50–73
[36] Xu W Y and Chamzas C 1983 On the extrapolation of band-limited functions with energy constraints *IEEE Trans. Acoustics, Speech Signal Processing* **31** 1222–34
[37] Zayed A I 1993 *Advances in Shannon's Sampling Theory* (Boca Raton, FL: Chemical Rubber Company)
[38] Zhou X W and Xia X-G 1986 On a conjecture of band-limited signal extrapolation *Kexue Tongbao (Chin. Sci. Bull.)* **31** 1593–7
[39] Zhou X W and Xia X-G 1989 A Sanz–Huang's conjecture on band-limited signal extrapolation with noises *IEEE Trans. Acoustics, Speech Signal Processing* **37** 1468–72
[40] Zhou X W and Xia X-G 1989 The extrapolations of high dimensional band-limited signals *IEEE Trans. Acoustics, Speech Signal Processing* **37** 1576–80

A Quantitative Analysis of SNR in the Short-Time Fourier Transform Domain for Multicomponent Signals

Xiang-Gen Xia

***Abstract*—A quantitative analysis is given for the signal-to-noise ratio (SNR) in the short-time Fourier transform domain for multicomponent signals in additive white noise. It is shown that the SNR is increased on the order of $O(N/K)$, where K is the number of components of a signal, N/T is the sampling rate, and T is the window size. The SNR increase rate is optimal for given K. For this result, the SNR definition is generalized, which is suitable for signals not only in the time domain but also in other domains. This theory is illustrated by one numerical example.**

I. Introduction

Time-frequency analysis [11]–[12] has become an important technique in analyzing wideband/nonstationary signals in various applications including inverse synthetic aperture radar (ISAR) imaging [1], biomedical signal analysis [2]–[3], speech signal analysis [4], and FM radio communications [5]. One of the most important features of this technique is that it usually increases the signal-to-noise ratio (SNR) in the joint time–frequency (TF) domain. This is particularly advantageous for signals that are difficult to detect in the time or frequency domain alone. The reason for this important feature can be stated as follows. A joint TF transform usually spreads noise from one dimension (the time or frequency) into two dimensions (the joint time and frequency) while it usually concentrates a signal in localized regions in the TF plane. A number of research results on the estimation of time-varying frequencies have appeared, such as [5]–[7] with Wigner–Ville distributions. However, there is little on quantitative analysis for the SNR increase for any joint TF transform, which is certainly an important issue in practical applications in signal detection by using thresholding.

In the conventional SNR definition, the mean power is taken over the whole domain of a signal. If the signal is stationary in this domain, this definition works fine. However, if the signal is not stationary in this domain, such as a single tone signal in the frequency domain, this definition is no longer suitable. In this correspondence, we first generalize the SNR definition so that it is not only suitable for signals in the time domain but also in other domains, such as the frequency domain and the joint TF domain. We then present a quantitative analysis of the SNR increase rate in the joint TF domain for the short-time Fourier transform with rectangular windows, where multicomponent signals in additive white noise are considered. The main result can be stated as follows.

- K number of monocomponents in a signal;
- T window size for the short-time Fourier transform;
- N/T sampling rate.

Manuscript received December 27, 1996; revised July 1, 1997. This work was supported in part by an initiative grant from the Department of Electrical and Computer Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377. The associate editor coordinating the review of this paper and approving it for publication was Dr. Akram Aldroubi.

The author is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: xxia@ee.udel.edu).

Publisher Item Identifier S 1053-587X(98)00508-X.

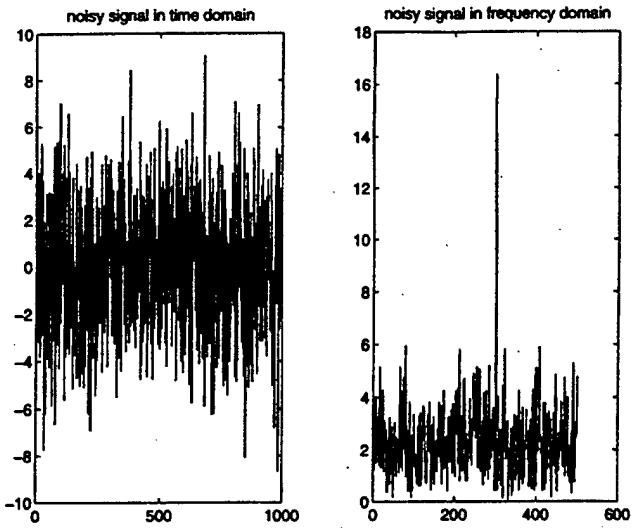

Fig. 1. Single tone signal.

N-point discrete Fourier transform is performed in each window. Then, the SNR in the joint TF domain is increased on the order of $O(N/K)$ when the window size T is small enough.

This correspondence is organized as follows. In Section II, we formulate a proper definition for SNR in different domains. In Section III, we present the proposed quantitative approach to analyze the SNR increase rate in the joint TF domain. A numerical example is presented in Section IV to illustrate the proposed approach.

II. SNR in Different Domains

The conventional signal-to-noise ratio (SNR) is defined as the ratio of the mean power of the signal over the mean power of the noise, where the mean is taken over the whole time domain. It is formulated as follows. Let $y[n]$ be a distorted signal

$$y[n] = x[n] + \eta[n], \quad 0 \le n \le N-1 \tag{2.1}$$

where $x[n]$ is a signal, and $\eta[n]$ is an additive white noise with variance σ^2. The SNR is defined as

$$\text{SNR} = \frac{\sum_{n=0}^{N-1} |x[n]|^2}{N\sigma^2}. \tag{2.2}$$

This SNR is used quite often in describing the noise level relative to the signal and in distinguishing the signal from noise in stationary environments. When the SNR is too low, in general, it is impossible to distinguish the signal $x[n]$ from $y[n]$. However, for some special kinds of signals $x[n]$, such as narrowband signals, it is possible to detect the signal in the Fourier transform domain, even when the SNR is of negative decibels. An example is shown in Fig. 1, where the SNR $= -11$ dB and the signal x is a single tone signal.

According to the SNR definition in (2.2), an orthogonal transform does not change the SNR, i.e., the SNR in the transform domain is exactly equal to the SNR in the time domain. This is because of the energy preservation property of orthogonal transforms. This implies that the SNR of the signal in the frequency domain in Fig. 1(b) is still -11 dB. However, one can clearly see the signal in the frequency domain. This suggests that the SNR definition in (2.2) is not proper to

judge the possibility of detecting the signal in the frequency domain in Fig. 1(b). It should not be surprising since the signal in Fig. 1(b) is not stationary, and the mean power over the whole frequency domain is, of course, not proper to the signal with a single spike.

The above observation suggests that the SNR definition is transform-domain dependent and should relate to the bandwidth of a signal occupied in that domain. We now introduce the following SNR definition in a domain.

Suppose the expression (2.1) is already in a transform domain, where n is the discrete variable in the transform domain. Assume the additive white noise $\eta[n]$ in (2.1) occupies the full band in the transform domain. For the signal $x[n]$ of length $N, 0 \le n \le N-1$, let

$$\mathcal{B} \triangleq \left\{ n : 0 \le n \le N-1 \text{ and } |x[n]|^2 \ge 0.5 \max_{0 \le n \le N-1} |x[n]|^2 \right\} \tag{2.3}$$

where the number 0.5 comes from the common 3-dB bandwidth definition in communications. Then, the SNR is defined as

$$\text{SNR} \triangleq \frac{\sum_{n \in \mathcal{B}} |x[n]|^2}{|\mathcal{B}|\sigma^2} \tag{2.4}$$

where $|\mathcal{B}|$ denotes the cardinality of the set $\mathcal{B}$. Notice that this definition is similar to the SNR definition in communications, where the signal is only considered in its bandwidth.

One can clearly see that the SNR in (2.4) is always greater than or equal to the SNR in (2.2) because the mean in (2.4) is only taken over the first large values in the whole domain. With the SNR definition in (2.4), the SNR in the time domain for the signal in Fig. 1(a) is -8.4 dB, but the SNR in the frequency domain for the signal in Fig. 1(b) is 16.3 dB. Although about 2.6 dB SNR is increased over the original definition in (2.2), the SNR in the frequency domain is significantly better than the old SNR, that is, -11 dB, in describing the signal characteristics over the noise. The time domain SNR increase is consistent for relatively stationary signals without dramatic jumps in the time domain.

III. SNR in the Joint TF Domain

In this section, we analyze the SNR in the joint TF domain for the short-time Fourier transform, where the SNR defined in (2.4) is used. In order to do so, we first describe a multicomponent signal model.

A. Multicomponent Signal Model

Throughout the rest of this paper, we use the following multicomponent signal model:

$$y(t) = \sum_{k=1}^{K} x_k(t) + \eta(t), \quad 0 \le t \le T_0 \tag{3.1}$$

where we have the following assumptions

1) t is the continuous-time variable and limited in the finite observation interval $[0, T_0]$.
2) $\eta(t)$ is an additive white noise process with mean 0 and variance σ^2. It is not differentiable at any time $t \in [0, T_0]$ and independent of $x_k(t), 1 \le k \le K$.
3) For each $k, 1 \le k \le K$, $x_k(t)$ is a monocomponent time-varying signal, i.e.,

$$x_k(t) = A_k(t) e^{j\phi_k(t)} \tag{3.2}$$

where $A_k(t)$ is the slowly varying amplitude envelope of $x_k(t)$, and $\phi_k(t)$ is the phase of $x_k(t)$. The magnitude of the first order derivative $A_k'(t)$ is upper bounded by A_k, i.e., $|A_k'(t)| \le A_k$ for a positive constant A_k, and the magnitude of the second-order derivative $\phi_k''(t)$ is also upper bounded by ϕ_k, i.e., $|\phi_k''(t)| \le \phi_k$ for a positive constant ϕ_k for all $t \in [0, T_0]$.
4) The K instantaneous frequencies $\phi_k'(t)$, $1 \le k \le K$, are distinct.

Additional details on multicomponent signals can be found in [8]. It can be easily shown that the process $y(t)$ in (3.1) has locally stationary behavior [9]–[10] in the following sense:

$$|R_{yy}(t+u, s+u) - R_{yy}(t, s)| \le C|u| \tag{3.3}$$

for a positive constant C, where R_{yy} denotes the autocorrelation function of $y(t)$.

As a remark, the nondifferentiability assumption 2) of $\eta(t)$ makes sense. An example of such processes is the Wiener process; see, for example, [13]. This assumption implies that any sampled segment of $\eta(t)$ in any time interval is a white noise and has flat Fourier spectrum.

B. Short-Time Fourier Transform for Multicomponent Signals and SNR Calculations

For each monocomponent signal $x_k(t)$ in (3.1), by 1)–3), it can be shown that there exists $\epsilon_k > 0$ such that for any $s \in (\epsilon_k, T_0 - \epsilon_k)$

$$x_k(s+t) \approx A_k(s) e^{j(\phi_k(s) + \phi_k'(s)t)}, \quad t \in [-\epsilon_k, \epsilon_k]$$

where the linear term $A_k'(s)t$ of t does not appear because of the "slowly varying" assumption in 3) on the amplitude envelope $A_k(t)$. Since we have only finite many monocomponent signals $x_k(t)$ in (3.1), there exists $\epsilon = \min\{\epsilon_k,\ 1 \le k \le K\} > 0$ such that for any $s \in (\epsilon, T_0 - \epsilon)$ and any $k, 1 \le k \le K$

$$x_k(s+t) \approx A_k(s) e^{j(\phi_k(s) + \phi_k'(s)t)}, \quad t \in [-\epsilon, \epsilon] \tag{3.4}$$

where ϵ depends on the constants T_0, A_k, ϕ_k, and $1 \le k \le K$.

With (3.4), at each time $s \in (\epsilon, T_0 - \epsilon)$, we apply N-point discrete Fourier transform (DFT) for the signal $y(t)$ for $t \in (s - \frac{T}{2}, s + \frac{T}{2}]$ with the sampling rate N/T for $T = 2\epsilon$. For convenience, we assume that N is even. The DFT is

$$P_y[m, l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} y\left((m+q)\frac{T}{N}\right) e^{-\frac{2\pi jql}{N}}, \quad 0 \le l \le N-1 \tag{3.5}$$

where m is in the range such that $(m - N/2 + 1)T/N \ge 0$, and $(m + N/2)T/N \le T_0$, i.e.,

$$\frac{N-2}{2} \le m \le \left(\frac{T_0}{T} - \frac{1}{2}\right) N.$$

The above P_y can be decomposed into

$$P_y[m, l] = \sum_{k=1}^{K} P_{x_k}[m, l] + P_\eta[m, l], \quad 0 \le l \le N-1, \ \frac{N-2}{2} \le m \le \left(\frac{T_0}{T} - \frac{1}{2}\right) N \tag{3.6}$$

where $P_{x_k}[m, l]$ and $P_\eta[m, l]$ are defined for $x_k(t)$ and $\eta(t)$:

$$P_{x_k}[m, l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} x_k\left((m+q)\frac{T}{N}\right) e^{-\frac{2\pi jql}{N}}, \quad 0 \le l \le N-1 \tag{3.7}$$

$$P_\eta[m, l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} \eta\left((m+q)\frac{T}{N}\right) e^{-\frac{2\pi jql}{N}}, \quad 0 \le l \le N-1. \tag{3.8}$$

Since $\eta(t)$ is a white noise process, for each m, the Fourier spectra $E(|P_\eta[m,l]|^2)$ are flat over the whole frequency domain $0 \le l \le N-1$, as mentioned in Section III-A. This implies that the mean power of the noise spectrum $P_\eta[m,l]$ is also σ^2, which is the same as in the time domain.

We next want to study the mean power of $P_{x_k}[m,l]$ for the signal. Using (3.4)

$$\begin{aligned} P_{x_k}[m,l] &\approx \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} A_k\left(m\frac{T}{N}\right) \\ &\quad \times e^{j\left\{\phi_k\left(m\frac{T}{N}\right)+\phi_k'\left(m\frac{T}{N}\right)(m+q)\frac{T}{N}-\frac{2\pi ql}{N}\right\}} \\ &= \frac{1}{\sqrt{N}} A_k\left(m\frac{T}{N}\right) e^{j\left\{\phi_k\left(m\frac{T}{N}\right)+\phi_k'\left(m\frac{T}{N}\right)m\frac{T}{N}\right\}} \\ &\quad \times \sum_{q=-N/2+1}^{N/2} e^{jq\frac{\phi_k'(m\frac{T}{N})T-2\pi l}{N}}. \end{aligned}$$

Therefore

$$|P_{x_k}[m,l]| \approx \left|A_k\left(m\frac{T}{N}\right)\right| \sqrt{N}\delta\left(l-\frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right). \quad (3.9)$$

By the assumption of distinct instantaneous frequencies $\phi_k'(m\frac{T}{N})$ for $1 \le k \le K$, the Fourier power spectra $|P_{x_k}[m,l]|^2$ are located at K different frequencies $\phi_k'(m\frac{T}{N})T/(2\pi), 1 \le k \le K$. This implies

$$\begin{aligned} &\left|\sum_{k=1}^{K} |P_{x_k}[m,l]|\right|^2 \\ &\quad \approx N\left|\sum_{k=1}^{K} A_k\left(m\frac{T}{N}\right)\delta\left(l-\frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right)\right|^2 \\ &\quad \approx N\sum_{k=1}^{K}\left|A_k\left(m\frac{T}{N}\right)\right|^2 \delta\left(l-\frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right). \end{aligned} \quad (3.10)$$

Therefore, for each fixed time $s = m\frac{T}{N}$, in the frequency domain

$$\max_{0\le l\le N-1}\left|\sum_{k=1}^{K}|P_{x_k}[m,l]|\right|^2 \ge N\sum_{k=1}^{K}\left|A_k\left(m\frac{T}{N}\right)\right|^2. \quad (3.11)$$

Now, let us come back to the time domain signal $y(m\frac{T}{N})$. The noise mean power is σ^2. The signal power at each time $t = m\frac{T}{N}$ is

$$\begin{aligned} \left|\sum_{k=1}^{K} x_k\left(m\frac{T}{N}\right)\right|^2 &\le \left(\sum_{k=1}^{K}\left|A_k\left(m\frac{T}{N}\right)\right|\right)^2 \\ &\le K\sum_{k=1}^{K}\left|A_k\left(m\frac{T}{N}\right)\right|^2. \end{aligned} \quad (3.12)$$

By comparing (3.11) with (3.12), it is clear that the following relationship between the SNR_{tf} in the joint TF domain of (3.6) and the SNR_t in the time domain of (3.1) at the sampling points $m\frac{T}{N}$:

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \ge 0.5\frac{N}{K} \quad (3.13)$$

where 0.5 comes from the SNR definition in (2.3)–(3.4). Therefore, as the window size T is small enough

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \ge O\left(\frac{N}{K}\right). \quad (3.14)$$

Notice that the assumption of small enough window size T is equivalent to the assumption of fast enough sampling rate N/T. The derivation of (3.14) implies the following theorem.

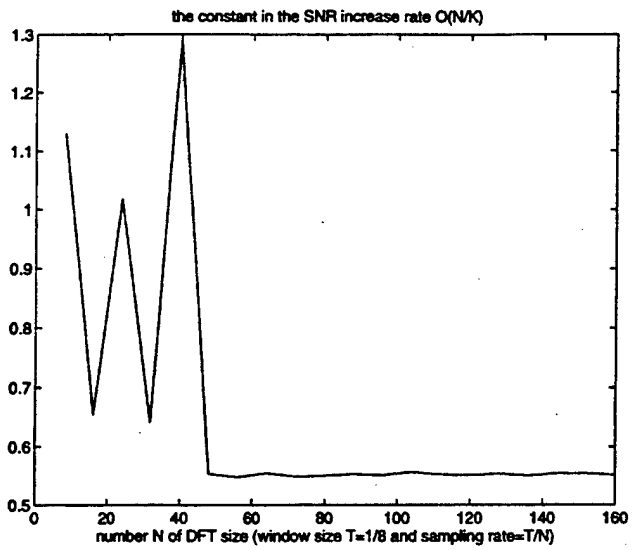

Fig. 2. SNR increase rate.

Theorem 1: For a multicomponent signal with K many monocomponents, the SNR in the joint TF domain with the short-time Fourier transform with the rectangular window of size T and the sampling rate N/T increases over the SNR in the time domain on the order of $O(N/K)$ when the sampling rate is fast enough. Given the number K, this increase rate $O(N/K)$ is optimal.

Proof: The first part has been proved by the above argument. The optimality can be proved by taking $A_k(t) = 1$ and $\phi_k(t) = c_k t^2$ for proper constants $c_k \ne 0$ for $1 \le k \le K$ and noticing that the inequalities in (3.9)–(3.12) become equalities in this case. □

IV. Numerical Example

For simplicity in computations, we choose the following two-component signal model:

$$y(t) = e^{j8\pi t^2} + e^{j\pi t^{2.5}} + \eta(t), \quad 0 \le t \le 2 \quad (4.1)$$

where $\eta(t)$ is an additive white Gaussian noise with mean 0 and variance $\sigma^2 = 9$. The window size for the short-time Fourier transform is 1/8. The following constant of the SNR increase rate in terms of the number of points N of the DFT is illustrated in Fig. 2:

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \Big/ \frac{N}{K}. \quad (4.2)$$

One can see that for this particular signal,

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \to 0.55\frac{N}{K}, \quad \text{as } N \to \infty. \quad (4.3)$$

From Fig. 2, one can also see that the constants of the SNR increase rate have large variance when the sampling rate is not large enough but almost become invariant when the sampling rate becomes large.

V. Conclusion

In this correspondence, we have quantitatively analyzed the SNR increase rate in the joint TF domain with the short-time Fourier transform over the SNR in the time domain for multicomponent signals in additive white noise. We have shown that the rate of the SNR increase is on the order of $O(N/K)$, where

DEPARTMENT OF THE NAVY
OFFICE OF NAVAL RESEARCH
ATLANTA REGIONAL OFFICE
ATLANTA FEDERAL CENTER
100 ALABAMA STREET SW STE 4R15
ATLANTA GA 30303-3104

4330
243-Atlanta:nvb
DLWL/F49620-97-1-0253
2 August 1999

From: Office of Naval Research, Atlanta Regional Office
To: AFOSR/NM, Attn: Dr. Jon A. Sjogren, 110 Duncan Avenue, Room B115, Bolling AFB, DC 20332-8050

SUBJ: **GRANT F49620-97-1-0253** WITH **UNIVERSITY OF DELAWARE (DLWL)**

1. This office is in the process of closing the subject grant. We have been advised that the final technical report has been submitted.

2. So that closeout may continue, please provide this office with certification of technical completion of the grant.

3. Please contact the undersigned by telephone at (404) 562-1616, fax: (404) 562-1610, and/or by E-mail{**bryantn@onr.navy.mil**} should additional information be required.

Natalie V. Bryant
NATALIE V. BRYANT
Grant Technician

DO NOT DETACH

FIRST ENDORSEMENT ON ONRRR/Atlanta ltr dtd

I certify that all technical requirements under this grant have been completed.

Jon A Sjogren 703 696 6564 AFOSR/NM
Scientific Officer

19 August 1999
Date

The Journal of Fourier Analysis and Applications
Volume 4, Issue 1, 1998

Smooth Local Sinusoidal Bases on Two-Dimensional L-Shaped Regions

Xiang-Gen Xia

Communicated by Pankaj N. Topiwala

ABSTRACT. In this article, we construct two-dimensional continuous/smooth local sinusoidal bases (also called Malvar wavelets) defined on L*-shaped regions. With this construction, one is able to construct local sinusoidal bases and lapped orthogonal transforms (LOT) on arbitrarily shaped regions. This work is motivated from and useful in object-based video coding, where a segmented moving object may have arbitrary shape and block transform coding of this object is needed.*

1. Introduction

It is known that, in block DCT transform coding, one first decomposes an image on a rectangular region into 8×8 or 16×16 blocks and then does 8×8 or 16×16 DCT on each small block. Due to the truncation of an image in the block decomposition, the blocking effects with block DCT degrade the performance in decoding at a low bit compression ratio. To eliminate the blocking effects, the lapped orthogonal transform (LOT) has been developed by Malvar et al. [15, 16], where overlaps between adjacent blocks in the decomposition are used. In LOT, a smooth transition of the DCTs between blocks is performed and therefore the blocking effects can be eliminated, while LOT does not increase the total number of pixels in the transform domain. For applications in image coding, see also [1].

Coifman and Meyer [8] generalized LOT from discrete-time signals to continuous-time waveforms with a general description for window functions. They constructed a family of smooth local sinusoidal bases, which are also called Malvar wavelets [18]. For more about local sinusoidal bases, see [1, 3, 4, 5, 8, 9, 10, 13, 17, 18, 19, 22, 23]. Further generalizations were made in recent literature,

Math Subject Classifications. 42A10, 42A16.
Keywords and Phrases. Local sinusoidal bases, Malvar wavelets, lapped orthogonal transform, object-based video coding.
Acknowledgements and Notes. His work was partially supported by an initiative grant from the Department of Electrical and Computer Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

ISSN 1069-5869

see for example [1, 3, 4, 11, 13, 14, 17, 19, 20, 21, 22, 23]. In particular, two dimensional nonseparable smooth local sinusoidal bases were constructed in [22] on rectangular regions and in [23] on hexagons. Discrete forms were discussed in [11, 14, 20, 21]. An important requirement for LOT is that the domain of an image must be rectangular. This, however, may not be true in object-based video coding [2, 6, 7, 12]. In object-based video coding, one usually first segments moving objects from image frames and then codes the motion vectors and the segmented moving objects. There are two intuitive ways to code a segmented object. One is to mask an object using a larger rectangular window that completely covers the object and do block DCT/LOT for the image on the masked rectangular region (see Fig. 1). This way usually wastes bits of coding when the shape of an object is not regular due to the inclusion of redundant areas (see Fig. 1). The other is to code the boundary of an object and code the content inside the boundary separately. Although this way does not include redundant area, it does require block DCT/LOT to be applicable to images defined on an arbitrarily shaped region.

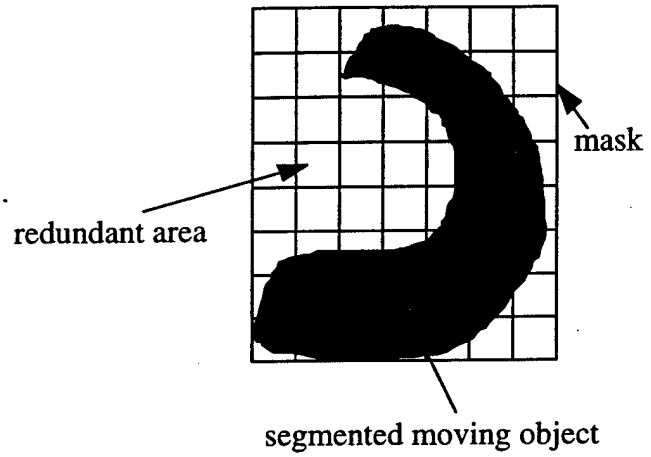

FIGURE 1. Segmented moving object and rectangular mask.

Another way to code a segmented moving object is between the above two ways, which masks the object by using small rectangular blocks [see Fig. 2(a)]. With this masking method, the redundant area is clearly smaller than the one in Fig. 1. The question now is whether LOT can be implemented for the rectangular blocks in Fig. 2(a). To study this question, we decompose the mask in Fig. 2(a) into two parts: a rectangularly shaped region part as shown in Fig. 2(b) and a nonrectangularly shaped boundary region part as shown in Fig. 2(c). For the rectangularly shaped region part, the standard constructions apply. Thus, the question is reduced into whether LOT can be implemented for the domain shown in Fig. 2(c). We call the regions with the shapes in Fig. 2(c) **L**-*shaped regions*. As long as LOT is implementable on **L**-shaped regions in Fig. 2(c), LOT is implementable on all masks shown in Fig. 2(a) which cover arbitrarily shaped regions. Based on this observation, in the rest of this article we focus on local sinusoidal bases/Malvar wavelets/LOT on **L**-shaped regions shown in Fig. 2(c).

This article is organized as follows. In Section 2, we construct continuous/smooth local sinusoidal bases on **L**-shaped regions. In particular, we present a set of conditions on two-dimensional window functions for two-dimensional continuous/smooth local sinusoidal bases on **L**-shaped regions. The conditions are different but similar to those in [8, 22, 23]. In Section 3, we present some numerical examples of both window and local sinusoidal bases on **L**-shaped regions. In Section 4, we briefly mention the construction of LOT on **L**-shaped regions, i.e., the discrete version of local sinusoidal bases on **L**-shaped regions. We also present an application of LOT on **L**-shaped regions

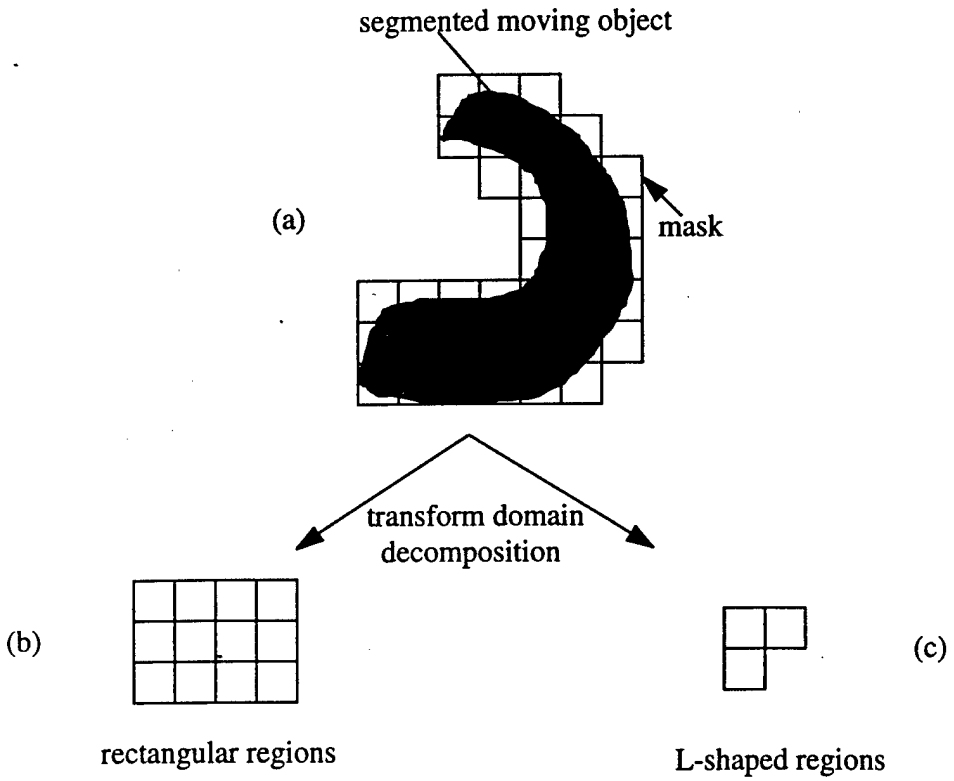

FIGURE 2. Segmented moving object and nonrectangular mask.

in image compression. In Section 5, we conclude by discussing local sinusoidal bases on mixed regions of L-shaped and rectangular regions.

2. A Theory for Local Sinusoidal Bases on L-Shaped Regions

In this section, we build a general theory for the construction of local sinusoidal bases on L-shaped regions. We first describe the problem precisely.

An L-shaped region A to work on in the following is shown in Fig. 3(a). The region A consists of three small rectangular regions: A_1, A_2, and A_3 as shown in Fig. 3(a). In what follows, small bold English letters, such as $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2)$ for real x_l, y_l, $l = 1, 2$, always denote two-dimensional vectors in $\mathbf{R}^2$. The goal of this article is to build smooth local sinusoidal bases/Malvar wavelets/LOT defined on the L-shaped region A from sinusoidal bases on A_j.

Let $f_{j,k}(\mathbf{x})$, $k \in \mathbf{Z}$, be an orthonormal basis defined on A_j for the signal space $L^2(A_j)$, $j = 1, 2, 3$, where $L^2(B)$ denotes all square integrable functions on the region B. A trivial method for forming an orthonormal basis for $L^2(A)$ is simply to use the truncation window $\chi_{A_j}(\mathbf{x})$, 1 for $\mathbf{x} \in A_j$ and 0 otherwise, and form $f_{j,k}(\mathbf{x})\chi_{A_j}(\mathbf{x})$, $j = 1, 2, 3$, $k \in \mathbf{Z}$, $\mathbf{x} \in A$. This is equivalent to using block DCT when $f_{j,k}$ are products of two cosine functions with discrete forms. Clearly, the basis elements $f_{j,k}(\mathbf{x})\chi_{A_j}(\mathbf{x})$ may have discontinuities which may cause blocking effects as discussed in the Introduction. The purpose of the rest of this article is to construct continuous/smooth basis for $L^2(A)$ from continuous/smooth local bases $f_{j,k}$ defined on A_j, $j = 1, 2, 3$. The basic idea to achieve the goal is similar to previous work: include overlaps between these three rectangular regions A_j and replace the truncation window χ_{A_j} with some better designed windows W_j. Before going to the details, we define some notations.

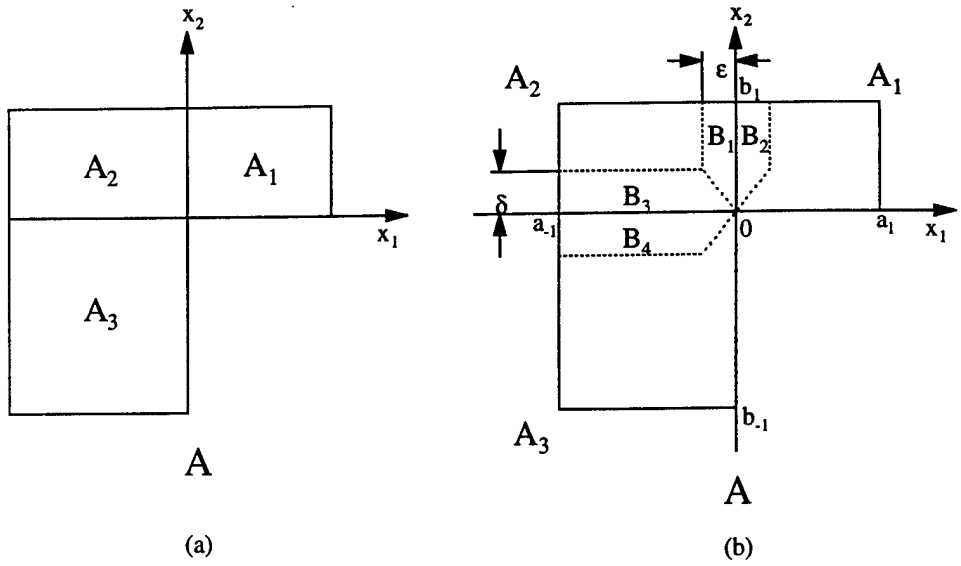

FIGURE 3. An L-shaped region A.

2.1 Notations

We divide the L-shaped region A in Fig. 3(a) as follows [see Fig 3(b)]:

$$\begin{aligned} A_1 &= \{\mathbf{x} = (x_1, x_2) : \ a_0 \le x_1 \le a_1, \ b_0 \le x_2 \le b_1\} \\ A_2 &= \{\mathbf{x} = (x_1, x_2) : \ a_{-1} \le x_1 \le a_0, \ b_0 \le x_2 \le b_1\} \\ A_3 &= \{\mathbf{x} = (x_1, x_2) : \ a_{-1} \le x_1 \le a_0, \ b_{-1} \le x_2 \le b_0\}, \end{aligned}$$

where $a_0 = b_0 = 0, a_{-1}, b_{-1} < 0$, and $a_1, b_1 > 0$.

The overlaps between the three rectangular regions A_j are shown in Fig. 3(b), which are bounded by the dotted lines. The symbols ϵ and δ denote the single width of the overlaps between A_1 and A_2 in the x_1 direction, and between A_2 and A_3 in the x_2 direction, respectively, shown in Fig. 3(b). The overlaps consist of four nonoverlapped regions $B_l, l = 1, 2, 3, 4$, shown in Fig. 3(b):

$$\begin{aligned} B_1 &= \{\mathbf{x} = (x_1, x_2) : \ -\epsilon \le x_1 \le a_0, \ \delta \le x_2 \le b_1\} \\ &\quad \cup \{\mathbf{x} = (x_1, x_2) : \ -\epsilon \le x_1 \le a_0, \ -\delta x_1/\epsilon \le x_2 \le \delta\} \\ B_2 &= \{\mathbf{x} = (x_1, x_2) : \ (-x_1, x_2) \in B_1\} \\ B_3 &= \{\mathbf{x} = (x_1, x_2) : \ a_{-1} \le x_1 \le -\epsilon, \ b_0 \le x_2 \le \delta\} \\ &\quad \cup \{\mathbf{x} = (x_1, x_2) : \ -\epsilon \le x_1 \le a_0, \ b_0 \le x_2 \le -\delta x_1/\epsilon\} \\ B_4 &= \{\mathbf{x} = (x_1, x_2) : \ (x_1, -x_2) \in B_3\} \end{aligned}$$

With these overlaps, the extended regions of A_j, denoted by $\tilde{A}_j$ shown in Fig. 4, $j = 1, 2, 3$, are: $\tilde{A}_1 = B_1 \cup A_1$, $\tilde{A}_2 = B_2 \cup B_4 \cup A_2$, and $\tilde{A}_3 = B_3 \cup A_3$.

2.2 Theory for Construction

In order to construct continuous/smooth orthonormal bases for $L^2(A)$ from local orthonormal bases $f_{j,k}(\mathbf{x})$, $k \in \mathbf{Z}$, for $L^2(A_j)$, $j = 1, 2, 3$, first we need to extend $f_{j,k}(\mathbf{x})$ from the rectangular regions A_j to the L-shaped region A and then construct window functions W_j to window the extended local bases. We denote the extended bases of $f_{j,k}$ as $\tilde{f}_{j,k}$, where odd and even extensions are also

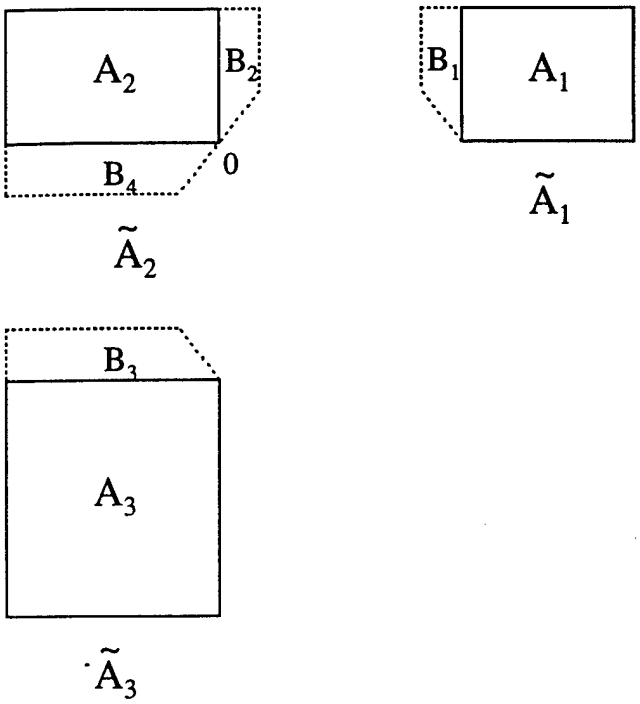

FIGURE 4. Extended regions of A_j.

used as follows.

$$\tilde{f}_{1,k}(\mathbf{x}) = \begin{cases} f_{1,k}(\mathbf{x}), & \mathbf{x} \in A_1, \\ f_{1,k}(-x_1, x_2), & \mathbf{x} = (x_1, x_2) \in B_1, \\ 0, & \mathbf{x} \in A \text{ but } \mathbf{x} \notin \tilde{A}_1, \end{cases} \tag{2.1}$$

$$\tilde{f}_{2,k}(\mathbf{x}) = \begin{cases} f_{2,k}(\mathbf{x}), & \mathbf{x} \in A_2, \\ -f_{2,k}(-x_1, x_2), & \mathbf{x} = (x_1, x_2) \in B_2, \\ -f_{2,k}(x_1, -x_2), & \mathbf{x} = (x_1, x_2) \in B_4, \\ 0, & \mathbf{x} \in A \text{ but } \mathbf{x} \notin \tilde{A}_2, \end{cases} \tag{2.2}$$

$$\tilde{f}_{3,k}(\mathbf{x}) = \begin{cases} f_{3,k}(\mathbf{x}), & \mathbf{x} \in A_3, \\ f_{3,k}(x_1, -x_2), & \mathbf{x} = (x_1, x_2) \in B_3, \\ 0, & \mathbf{x} \in A \text{ but } \mathbf{x} \notin \tilde{A}_3, \end{cases} \tag{2.3}$$

where $k \in \mathbf{Z}$. Clearly, $\tilde{f}_{j,k}(\mathbf{x})$. $k \in \mathbf{Z}$ are supported on the extended region $\tilde{A}_j$ of A_j, $j = 1, 2, 3$.

The conditions on window functions W_j are the following.

(a) $W_1(\mathbf{x}) = 1$ for $\mathbf{x} \in A_1$ but $\mathbf{x} \notin B_2$,
$W_2(\mathbf{x}) = 1$ for $\mathbf{x} \in A_2$ but $\mathbf{x} \notin B_1 \cup B_3$,
$W_3(\mathbf{x}) = 1$ for $\mathbf{x} \in A_3$ but $\mathbf{x} \notin B_4$.

(b) $W_j(\mathbf{x}) = 0$ for $\mathbf{x} \in A$ but $\mathbf{x} \notin \tilde{A}_j$ for $j = 1, 2, 3$.

(c) $W_1(x_1, x_2) = W_2(-x_1, x_2)$ for $\mathbf{x} = (x_1, x_2) \in B_1 \cup B_2$,
$W_2(x_1, x_2) = W_3(x_1, -x_2)$ for $\mathbf{x} = (x_1, x_2) \in B_3 \cup B_4$.

(d) $W_1^2(\mathbf{x}) + W_2^2(\mathbf{x}) = 1$ for $\mathbf{x} \in B_1 \cup B_2$,
$W_2^2(\mathbf{x}) + W_3^2(\mathbf{x}) = 1$ for $\mathbf{x} \in B_3 \cup B_4$.

The support of W_j is also $\tilde{A}_j$ for $j = 1, 2, 3$.

Finally, we form

$$u_{j,k}(\mathbf{x}) = \tilde{f}_{j,k}(\mathbf{x})W_j(\mathbf{x}),\ \mathbf{x} \in A,\ j = 1, 2, 3, k \in \mathbf{Z}, \tag{2.4}$$

and have the following theorem.

Theorem 1.

The functions $u_{j,k}(\mathbf{x})$ for $j = 1, 2, 3$, $k \in \mathbf{Z}$, and $\mathbf{x} \in A$, form an orthonormal basis for $L^2(A)$ when $f_{j,k}(\mathbf{x})$, $k \in \mathbf{Z}$, $\mathbf{x} \in A_j$, form an orthonormal basis for $L^2(A_j)$, $j = 1, 2, 3$.

Proof. The proofs of the orthogonality and the completeness of $u_{j,k}$ are similar to the ones for Malvar wavelets on rectangular regions studied in [22]. The details are omitted here. □

Remark. The odd and even extensions in (2.1) through (2.3) are also called the folding processes in [1, 3, 4, 13, 17, 19]. The window functions in (a) through (d) also can be characterized similar to the one-dimensional case studied in [4]. □

3. A Family of Continuous/Smooth Local Sinusoidal Bases on L-Shaped Regions

With the general theory in Section 2.2 we want to construct a family of continuous/smooth local sinusoidal bases on L-shaped regions. The local bases $f_{j,k}$ are separable sine bases as follows.

$$\begin{aligned} f_{1,k_1,k_2}(x_1, x_2) &= \frac{2}{\sqrt{(a_1 - a_0)(b_1 - b_0)}} \sin\left(\pi\left(k_1 + \frac{1}{2}\right)\frac{x_1 - a_1}{a_0 - a_1}\right) \\ &\quad \sin\left(\pi\left(k_2 + \frac{1}{2}\right)\frac{x_2 - b_0}{b_1 - b_0}\right), \end{aligned} \tag{3.1}$$

$$\begin{aligned} f_{2,k_1,k_2}(x_1, x_2) &= \frac{2}{\sqrt{(a_0 - a_{-1})(b_1 - b_0)}} \sin\left(\pi\left(k_1 + \frac{1}{2}\right)\frac{x_1 - a_0}{a_{-1} - a_0}\right) \\ &\quad \sin\left(\pi\left(k_2 + \frac{1}{2}\right)\frac{x_2 - b_0}{b_1 - b_0}\right), \end{aligned} \tag{3.2}$$

$$\begin{aligned} f_{3,k_1,k_2}(x_1, x_2) &= \frac{2}{\sqrt{(a_0 - a_{-1})(b_0 - b_{-1})}} \sin\left(\pi\left(k_1 + \frac{1}{2}\right)\frac{x_1 - a_0}{a_{-1} - a_0}\right) \\ &\quad \sin\left(\pi\left(k_2 + \frac{1}{2}\right)\frac{x_2 - b_{-1}}{b_0 - b_{-1}}\right), \end{aligned} \tag{3.3}$$

where $k_1, k_2 = 0, 1, 2, \ldots$.

Next, we construct window functions $W_j(\mathbf{x})$. The idea for the construction is the following. We draw L-shaped lines in an L-shaped region A, which are parallel to the boundary of A (see Fig. 5). We treat these lines as one-dimensional domains where one-dimensional window functions are defined. The overlaps for these one-dimensional window functions are the intervals bounded by the dots, i.e., the intersections of the L-shaped lines with regions $B_1 \cup B_2$ and $B_3 \cup B_4$. There are two kinds of such L-shaped lines. For the first one, the overlaps are separated and for the second one, the overlaps are adjacent but do not intersect except at the boundaries (see Fig. 5). Notice that these lines are not closed, which is not like the case for hexagons studied in [23].

With the above idea, the following construction for window functions W_j follows.

Step 1. 1-regions.

$$W_1(\mathbf{x}) = 1 \text{ for } \mathbf{x} \in A_1 \text{ but } \mathbf{x} \notin B_2,$$

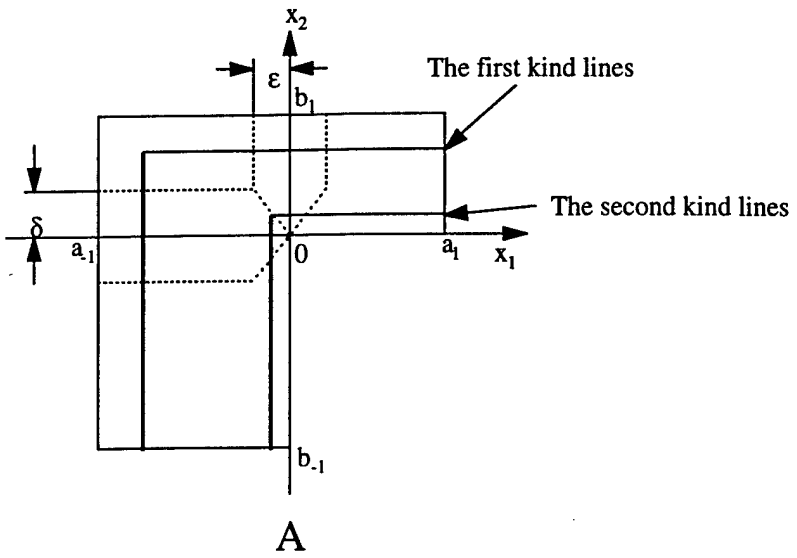

FIGURE 5. Window functions constructed from one-dimensional window functions.

$$W_2(\mathbf{x}) = 1 \text{ for } \mathbf{x} \in A_2 \text{ but } \mathbf{x} \notin B_1 \cup B_3 ,$$
$$W_3(\mathbf{x}) = 1 \text{ for } \mathbf{x} \in A_3 \text{ but } \mathbf{x} \notin B_4 .$$

Step 2. 0-regions.

$$W_j(\mathbf{x}) = 0 \text{ for } \mathbf{x} \in A \text{ but } \mathbf{x} \notin \tilde{A}_j,\ j = 1, 2, 3 .$$

Step 3. Overlapped regions.

On the region $B_1 \cup B_2$,

$$W_1(\mathbf{x}) = W_1(x_1, x_2) = \begin{cases} \sin\left(\frac{\pi}{4\epsilon}\{x_1 - a_0 + \epsilon\}\right), & |x_1| \le \epsilon, \delta \le x_2 \le b_1 , \\ \sin\left(\frac{\pi\delta}{4x_2\epsilon}\{x_1 - a_0 + \frac{x_2}{\delta}\epsilon\}\right), & |x_1| \le \frac{x_2}{\delta}\epsilon, b_0 \le x_2 \le \delta, \end{cases}$$

$$W_2(\mathbf{x}) = W_2(x_1, x_2) = \begin{cases} \cos\left(\frac{\pi}{4\epsilon}\{x_1 - a_0 + \epsilon\}\right), & |x_1| \le \epsilon, \delta \le x_2 \le b_1 , \\ \cos\left(\frac{\pi\delta}{4x_2\epsilon}\{x_1 - a_0 + \frac{x_2}{\delta}\epsilon\}\right), & |x_1| \le \frac{x_2}{\delta}\epsilon, b_0 \le x_2 \le \delta . \end{cases}$$

On the region $B_3 \cup B_4$,

$$W_2(\mathbf{x}) = W_2(x_1, x_2) = \begin{cases} \sin\left(\frac{\pi}{4\delta}\{x_2 - b_0 + \delta\}\right), & |x_2| \le \delta, a_{-1} \le x_1 \le -\epsilon, \\ \sin\left(\frac{-\pi\epsilon}{4x_1\delta}\{x_2 - b_0 + \frac{-x_1}{\epsilon}\delta\}\right), & |x_2| \le \frac{-x_1}{\epsilon}\delta, -\epsilon \le x_1 \le a_0 , \end{cases}$$

$$W_3(\mathbf{x}) = W_3(x_1, x_2) = \begin{cases} \cos\left(\frac{\pi}{4\delta}\{x_2 - b_0 + \delta\}\right), & |x_2| \le \delta, a_{-1} \le x_1 \le -\epsilon, \\ \cos\left(\frac{-\pi\epsilon}{4x_1\delta}\{x_2 - b_0 + \frac{-x_1}{\epsilon}\delta\}\right), & |x_2| \le \frac{-x_1}{\epsilon}\delta, -\epsilon \le x_1 \le a_0 . \end{cases}$$

For general one-dimensional window functions $W_j(x)$, the construction of $W_j(\mathbf{x})$ on the regions $B_1 \cup B_2$ and $B_3 \cup B_4$ can be obtained by replacing the above sin and cos with W_j properly. Notice that the above $W_j(\mathbf{x})$ are continuous everywhere else inside A but at the origin and are basically generated from one-dimensional window functions for smooth local sinusoidal bases. With the construction of smooth one-dimensional window functions $W_j(x)$, it is possible to construct smooth $W_j(\mathbf{x})$ in the sense of continuous $\partial^{k_1+k_2} W_j(x_1, x_2)/\partial x_1^{k_1} \partial x_2^{k_2}$ for some nonnegative integers k_1 and k_2 when $(x_1, x_2) \neq (0, 0)$.

Extend the local bases f_{j,k_1,k_2} in (3.1) through (3.3) from the domains A_j to A according to the extension method in (2.1) through (2.3), which are denoted by $\tilde{f}_{j,k_1,k_2}$. Then, form $u_{j,k_1,k_2}(\mathbf{x}) = \tilde{f}_{j,k_1,k_2}(\mathbf{x})W_j(\mathbf{x})$ for $j = 1, 2, 3$ and $k_1, k_2 = 0, 1, 2, \ldots$. Since functions $f_{j,k_1,k_2}(\mathbf{x})$ are zero at the origin, the windowed basis elements $u_{j,k_1,k_2}(\mathbf{x})$ are continuous in the L-shaped region A. The above window functions W_j clearly satisfy conditions (a) through (d). This proves the following result.

Theorem 2.

The above constructed functions $u_{j,k_1,k_2}(\mathbf{x})$, $j = 1, 2, 3$ and $k_1, k_2 = 0, 1, 2, \ldots$, are continuous and form an orthonormal basis for $L^2(A)$.

Next we want to see some numerical examples. Let $a_{-1} = -2$, $a_0 = 0$, $a_1 = 1$, $b_{-1} = -3$, $b_0 = 0$, $b_1 = 2$, $\epsilon = 0.5$, and $\delta = 1$. Figures 6 through 8 show the window functions $W_j(\mathbf{x})$ for $j = 1, 2, 3$, respectively. Figures 9 through 11 show the basis elements $u_{j,1,1}(\mathbf{x})$ for $j = 1, 2, 3$, respectively.

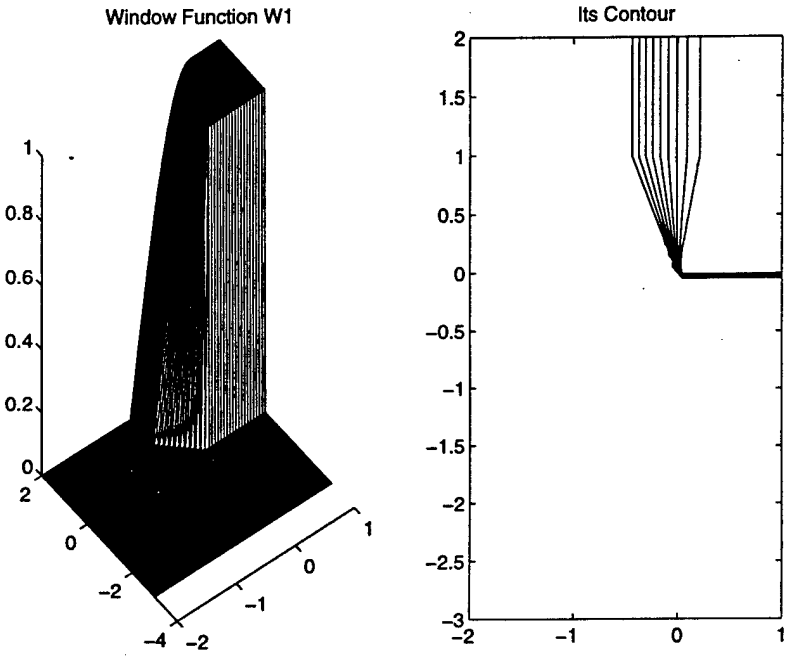

FIGURE 6. Window function $W_1(\mathbf{x})$ on A and its contour.

4. LOT on L-Shaped Regions and Application in Image Compression

In this section, we want to briefly introduce the construction of LOT on L-shaped regions. Then, we show a numerical example that shows that the SNR of the LOT on L-shaped regions performs better than the one of the LOT on rectangular regions or the block DCT.

The main difference between discrete-time and continuous-time local sinusoidal bases is that the variables x_1 and x_2 are integers and the overlap sizes ϵ and δ are also integers. As an example of constructions of discrete-time local sinusoidal bases or LOT on L-shaped regions, we consider

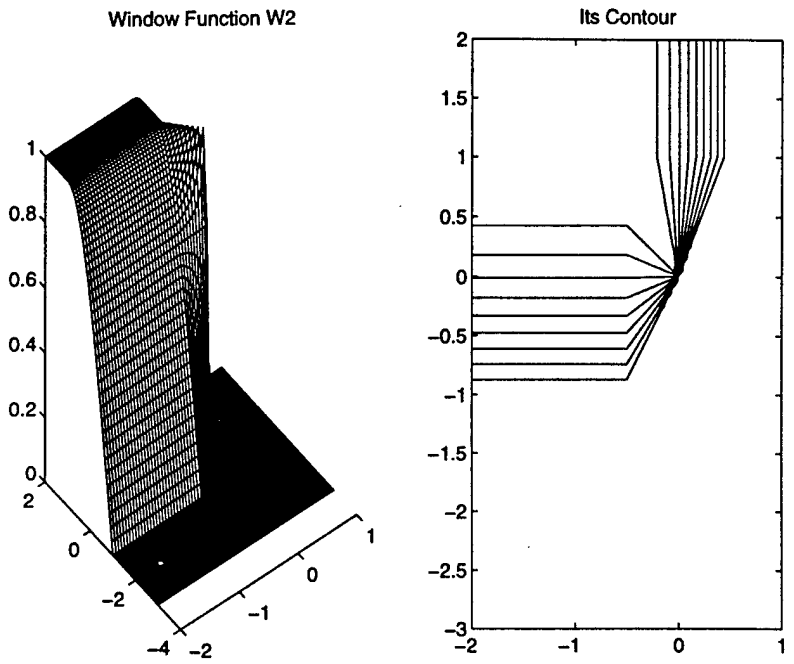

FIGURE 7. Window function $W_2(\mathbf{x})$ on A and its contour.

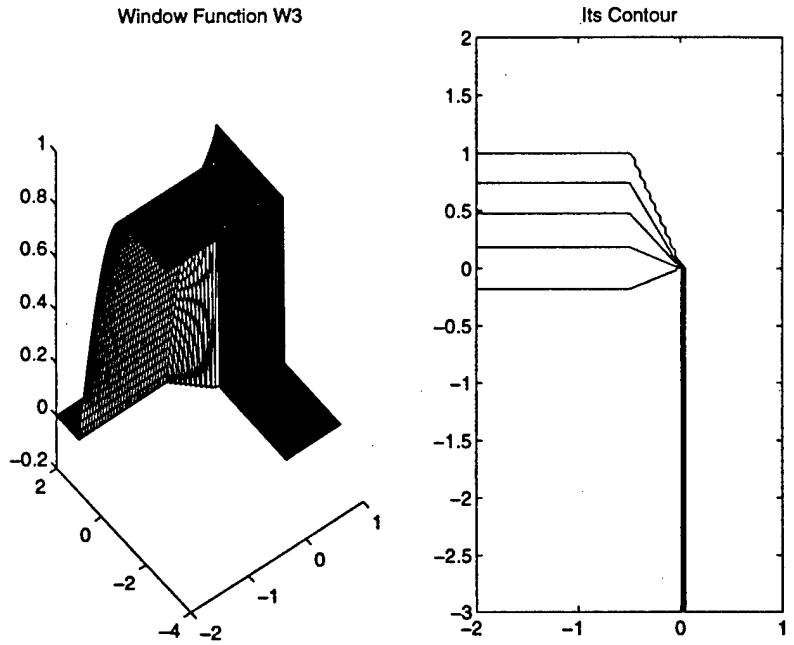

FIGURE 8. Window function $W_3(\mathbf{x})$ on A and its contour.

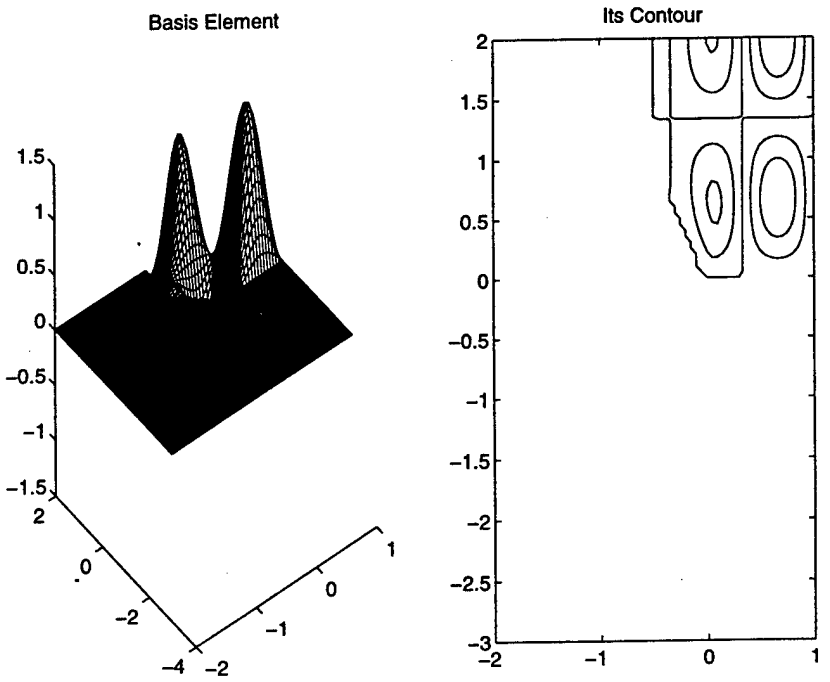

FIGURE 9. Basis $u_{1,1,1}(\mathbf{x})$ on A and its contour.

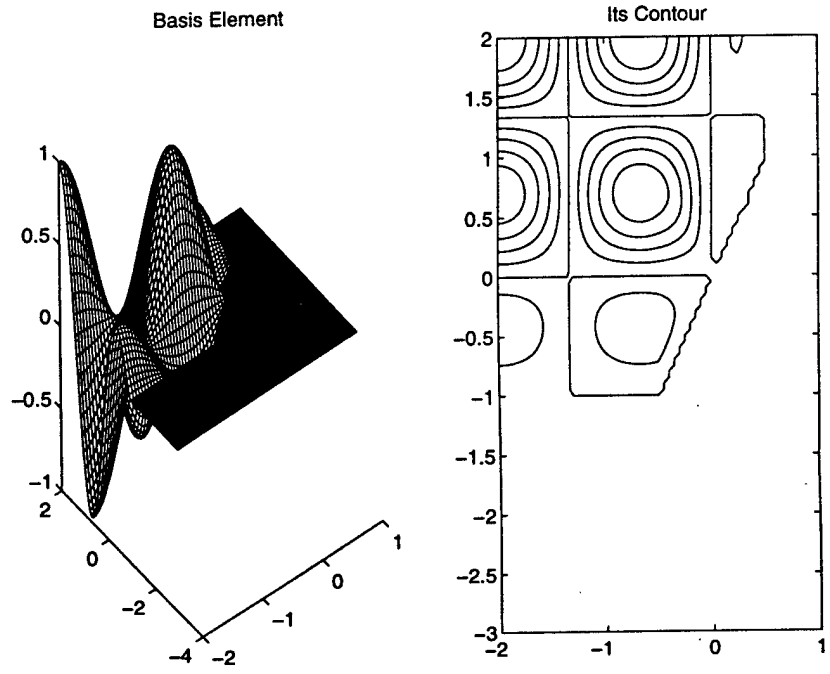

FIGURE 10. Basis $u_{2,1,1}(\mathbf{x})$ on A and its contour.

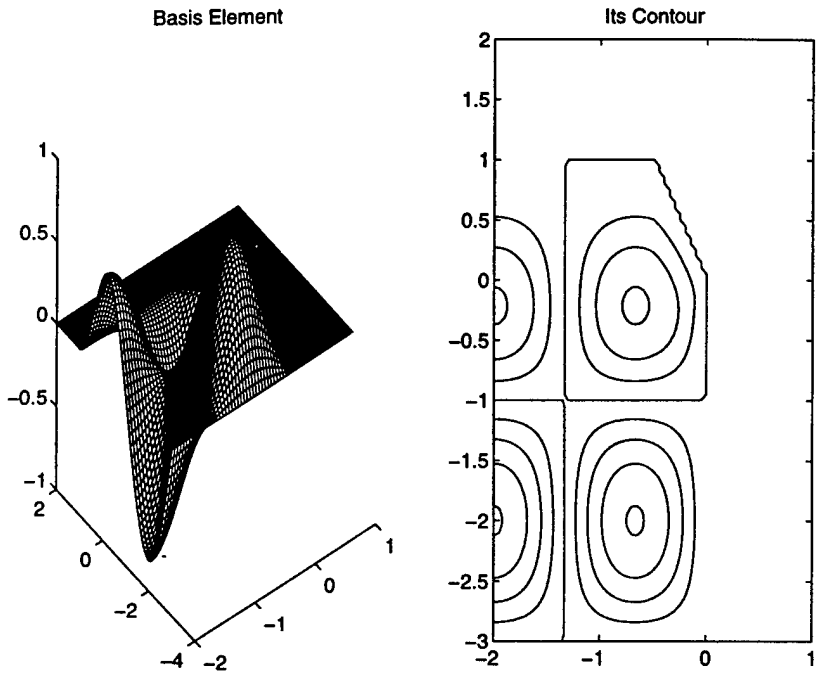

FIGURE 11. Basis $u_{3,1,1}(\mathbf{x})$ on A and its contour.

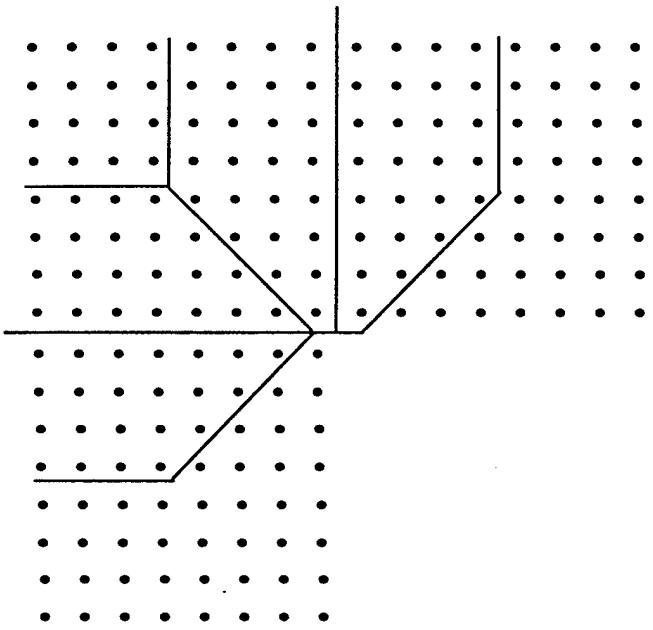

FIGURE 12. Discrete L-shaped region and the overlaps.

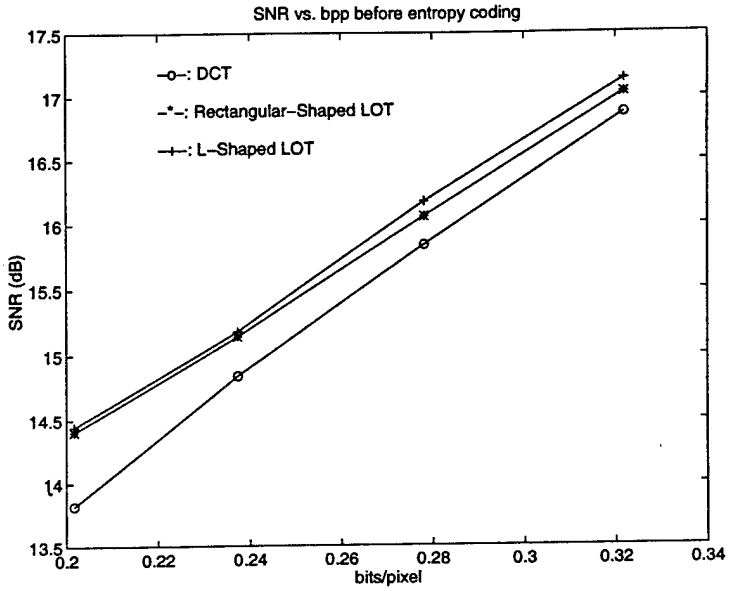

FIGURE 13. SNR performance comparison.

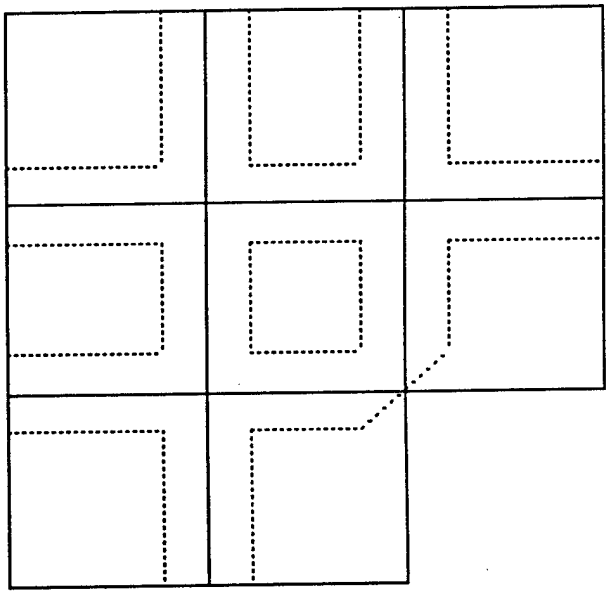

FIGURE 14. Mixed regions of L-shaped and rectangular regions.

three 8×8 blocks shown in Fig. 12. The overlap sizes are $\epsilon = \delta = 4$. Notice that, unlike the continuous-time case, the intersection set between $B_1 \cup B_2$ and $B_3 \cup B_4$ in the discrete-time case is empty. The overlaped regions are shown in Fig. 12, too. The rest is similar to the continuous-time case by replacing the continuous variables $x_i = n_i + 1/2$. For more details, see [20, 21, 22].

We have implemented a numerical example on applications of LOT on L-shaped regions in image compression. The test image has size 96×96 that is chosen for the convenience of the blocking. The signal-to-noise (SNR) ratio curves are shown in Fig. 13. One can clearly see the improvement of the LOT on L-shaped regions over the LOT on rectangular regions.

5. Conclusions

In this article, we have constructed continuous/smooth local sinusoidal bases/Malvar wavelets on L-shaped regions, which is motivated from object-based video coding. It is not hard to generalize the construction for mixed regions of rectangular and L-shaped regions, such as Fig. 14 with solid lines. An important point for the construction is the design of overlaps. For the region shown in Fig. 14, an overlap design is also shown with dotted lines. In this article, we use the construction for L-shaped regions and the separable construction or nonseparable construction studied in [22] for rectangular regions. Notice that the construction of continuous/smooth local sinusoidal bases at the most outside boundaries does not include any overlaps. This approach has been recently used in [24] for image compression.

Acknowledgments

The author would like to thank Bernie Soffer and Roy Matic at Hughes Research Laboratories for their encouragement and support, Roy Matic and David Shu for useful discussions in object-based video coding, and Yuri Owechiko and Yang Chen for their suggestions on the name of L-shaped regions. He also wishes to thank the reviewers for their careful reading of this manuscript and many useful comments that clarify the manuscript.

References

[1] Aharoni, G., Averbuch, A., Coifman, R., and Israeli, M. (1993). Local cosine transforms — a method for the reduction of the blocking effect in JPEG, *J. Math. Imag. Vision*, 3, 7–38.

[2] Apostolopoulos, J. and Lim, J. (1995). Coding of arbitrarily-shaped regions, *Visual Communications and Image Processing*, 1713–1726.

[3] Auscher, P. (1994). Remarks on the local Fourier bases, in *Wavelets: Mathematics and Applications*, J. Benedetto and M. Frazier, Eds., CRC Press, Boca Raton, FL, 204–218.

[4] Auscher, P., Weiss, G., and Wickerhauser, M.V. (1992). Local sine and cosine basis of Coifman and Meyer and the construction of smooth wavelets, in *Wavelets: A Tutorial in Theory and Applications*, C.K. Chui, Ed., Academic Press, Boston, 237–256.

[5] Bonami, A., Soria, F., and Weiss, G. (1993). Band-limited wavelets, *J. Geometric Anal.*, 3, 543–578.

[6] Chang, S.-F. and Messerschmitt, D. (1993). Transform coding of arbitrarily shaped image segments, *ACM Multimedia*, 83–90.

[7] Chen, H., Civanlar, M., and Haskell, B. (1994). A block transform coder for arbitrarily shaped image segments, *Intl. Conf. Image Processing*, 1, 85–89.

[8] Coifman, R. and Meyer, Y. (1991). Remarques sur l'analyse de Fourier à fenêt re, *C. R. Acad. Sci. Paris*, 312, Serie I, 259–261.

[9] Daubechies, I. (1994). Two recent results on wavelets: Wavelet bases for the interval, and biorthogonal wavelets diagonalizing the derivative operator, *Topics in the Theory and Applications of Wavelets,* L. Schumaker and G. Webb, Eds., Academic Press, Boston, MA, 237–248.

[10] Daubechies, I., Jaffard, S., and Journé, J.-L. (1991). A simple Wilson orthonormal basis with exponential decay, *SIAM J. Math. Anal.,* **22**, 554–572.

[11] de Queiroz, R.L. and Rao, K.R. (1993). Time-varying lapped transforms and wavelet packets, *IEEE Trans. Signal Processing,* Special Issue on Wavelets and Signal Processing, **41**, 3293–3305.

[12] Gile, M., Engelhardt, T., and Mehlan, R. (1989). Coding of arbitrarily shaped image segments based on a generalized orthogonal transform, *Signal Processing: Image Communication,* **1**, 153–180.

[13] Jawerth, B. and Sweldens, W. (1995). Biorthogonal smooth local trigonometric bases, *J. Fourier Anal. Appl.,* **2**, 109–132.

[14] Kovacevic, J. (1995). Local cosine bases in two dimensions, *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing,* **IV**, 2125–2128.

[15] Malvar, H.S. (1990). Lapped transforms for efficient transform subband coding, *IEEE Trans. on Acoustics, Speech, and Signal Processing,* **38**, 969–978.

[16] Malvar H.S. and Staelin, D.H. (1989). The LOT: Transform coding without blocking effects, *IEEE Trans. on Acoustics, Speech, and Signal Processing,* **37**, 553–559.

[17] Matviyenko, G. (1994). Optimized local trigonometric bases, Technical Reprot YALEU/DCS/RR-1041, Department of Computer Science, Yale University, New Haven, CT.

[18] Meyer, Y. (1993). Wavelets: Algorithms and applications, *SIAM,* Philadelphia.

[19] Wickerhauser, M.V. (1993). Smooth localized orthonormal bases, *C. R. Acad. Sci. Paris,* 316, Serie I, 423–427.

[20] Xia, X.-G. and Suter, B.W. (1994). Construction of perfect reconstruction time-varying FIR multirate filter banks with overlaps, *Proc. IEEE-SP Intl. Symp. Time-Frequency and Time-Scale Analysis,* Philadelphia.

[21] Xia, X.-G. and Suter, B.W. (1994). A systematic construction method for spatial-varying FIR filter banks with perfect reconstruction, *Proc. First IEEE Intl. Conf. Image Processing,* Austin, Texas.

[22] Xia, X.-G. and Suter, B.W. (1995). A family of two-dimensional nonseparable Malvar wavelets, *Appl. Computational Harmonic Anal.,* **2**, 243–256.

[23] Xia, X.-G. and Suter, B.W. (1996). Construction of Malvar wavelets on hexagons, *Appl. Computational Harmonic Anal.,* **3**, 65–71.

[24] Xia, X.-G. and Matic, R.M. (1996). Lapped orthogonal transforms on L-shaped regions and applications in object-based video coding, Patent pending, Hughes Research Laboratories.

Received February 5, 1996

Department of Electrical and Computer Engineering,
University of Delaware, Newark, DE 19716

A New Prefilter Design for Discrete Multiwavelet Transforms

Xiang-Gen Xia, *Member, IEEE*

Abstract— **In conventional wavelet transforms, prefiltering is not necessary due to the lowpass property of a scaling function. This is no longer true for multiwavelet transforms. A few research papers on the design of prefilters have appeared recently, but the existing prefilters are usually not orthogonal, which often causes problems in coding. Moreover, the condition on the prefilters was imposed based on the first-step discrete multiwavelet decomposition. In this paper, we propose a new prefilter design that combines the ideas of the conventional wavelet transforms and multiwavelet transforms. The prefilters are orthogonal but nonmaximally decimated. They are derived from a very natural calculation of multiwavelet transform coefficients. In this new prefilter design, multiple step discrete multiwavelet decomposition is taken into account. Our numerical examples (by taking care of the redundant prefiltering) indicate that the energy compaction ratio with the Geronimo–Hardin–Massopust 2 wavelet transform and our new prefiltering is better than the one with Daubechies D_4 wavelet transform.**

I. Introduction

NOW THAT single wavelet transforms are well-understood, multiwavelets recently have attracted much attention in the research community; see, for example, [1]–[20], [26]–[32], where several wavelet functions and scaling functions are used to expand a signal. The multiwavelet functions constructed by Geronimo *et al.* [2]–[4] have more desired properties than any single wavelet function, such as short support, symmetry, and smoothness. Although, in theory, they look more attractive than single wavelets, not much more advantages in practical applications over single wavelets have been found so far. In this author's opinion, the main reason behind this fact might be because of their improper discrete implementations. For single wavelet transforms, the discrete implementation automatically follows from their multiresolution structure, i.e., tree-structured two-channel filterbanks. In the tree-structured filterbank, lowpass and highpass filters are explicitly used, which is tight with the lowpass and the bandpass properties of the scaling and wavelet functions, respectively. Although, for multiwavelet transforms, the discrete implementation also follows from their multiresolution structure, the tree-structured filterbank becomes a tree-structured vector filterbank [1], [8] (or time-variant filterbank [13]). For a tree-structured vector filterbank, the lowpass and the highpass properties for the two vector filters are not as clear as those for the two filters in single wavelet transforms. It has been found in [1], [16]–[17] that in order to have a reasonable decomposition for discrete multiwavelet transforms, prefiltering is necessary. A prefilter design method was introduced in [1], [16]–[17], where the idea is based on the computability of the multiwavelet transform coefficients from uniformly sampled signals. Moreover, an interpretation of the "lowpass" and "highpass" properties for vector filters was introduced in [1] for the prefilter design criterion. The criterion is, however, only good for the first step discrete multiwavelet transform decomposition. The prefilters designed with this method may be nonorthogonal, which might kill the gain of the energy compaction in the transform domain after the decoding is performed. In [31], a different approach was proposed for perserving the orthogonality by using the approximation order criterion. In [32], balanced multiwavelets were studied, where prefiltering for these kinds of multiwavelets is not necessary, but other properties, such as the short supportness and the smoothness, are not as good as the GHM multiwavelets. Notice that in [1] and [8], it was also mentioned that when the "lowpass" filter $\mathbf{H}(\omega)$ satisfies $\mathbf{H}(0) = I$, prefiltering is not necessary.

In this paper, we introduce a new prefilter design by combining ideas in single wavelet transforms and multiwavelet transforms as follows. We first construct a function $\phi(t)$ with the lowpass property, i.e., its Fourier transform $\hat{\phi}(\omega)$ is 1 at $\omega = 0$, or $\hat{\phi}(0) = 1$, from the multiscaling functions and their translations such that $\phi(t-n)$, $n \in \mathbf{Z}$ form an orthonormal set. Notice that the function ϕ does not have to be a scaling function since the nested property is not required, i.e., a dilation equation may not be satisfied. Due to the lowpass property, a signal $f(t)$ can be well approximated by a linear combination of $2^{J/2}\phi(2^J t - n)$, $n \in \mathbf{Z}$ for a large J; meanwhile, $f(t)$ can also be well approximated by a linear combination of the multiscaling functions and their translations due to their multiresolution approximation property. Because of the lowpass property of ϕ and the orthogonality of $\phi(t-n)$, the coefficients in the linear combination of $2^{J/2}\phi(2^J t - n)$, $n \in \mathbf{Z}$ are proportational to $f(n/2^J)$; see, for example, [23]–[25], and [35]. The conversion between these two approximations naturally suggests a prefiltering for computing the multiwavelet transform coefficients at the highest resolution (or called approximation coefficients) from the samples $f(n/2^J)$ of the signal f. Then, the rest of

Manuscript received September 12, 1996; revised January 27, 1998. This work was supported in part by an initiative grant from the Department of Electrical and Computer Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377. The associate editor coordinating the review of this paper and approving it for publication was Dr. Truong Q. Nguyen.

The author is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: xxia@ee.udel.edu).

Publisher Item Identifier S 1053-587X(98)03918-X.

the multiwavelet transform coefficients (the lowest resolution coefficients and the detailed coefficients) follows from a tree-structured vector filterbank [1], [8]. We will see later that the lowpass condition imposed on the function ϕ is strongly related to the lowpass condition imposed on the combined filters of the prefilters and the multiscaling functions, which also relates to the one imposed on the combined filters of the prefilters and the cascaded vector filterbanks, i.e., multiple steps of the discrete multiwavelet transform decompositions. Notice that the above prefilter structure was first used in [30], but neither the *lowpass condition* on the function ϕ nor any *rationale* for introducing such ϕ was mentioned. Instead, in [30], signal-dependent optimal prefilters, in terms of the energy compaction criterion, were designed. The drawbacks are 1) that the computational load is high and 2) the signal dependency. In this paper, we systematically study the prefilter structure and its rationale. The prefilters are signal independent and orthogonal, and they only depend on multiwavelets.

II. Approximation of Lowpass Functions Using Multiscaling Functions and New Prefilter Structure

In this section, we want to motivate a new prefiltering for multiwavelet transform coefficient computation by approximating a lowpass function using multiscaling functions. To do so, let us first briefly review multiwavelets and matrix dilation equations. For more details about multiwavelets, see, for example, [1]–[20] and [26]–[32].

Consider N compactly supported scaling functions $\phi_l(t)$, $l = 1, 2, \ldots, N$ and their corresponding N mother wavelet functions $\psi_l(t)$, $l = 1, 2, \ldots, N$, where all the translations $\phi_l(t-k)$, $k \in \mathbf{Z}$, $l = 1, 2, \ldots, N$ are mutually orthogonal, and $\psi_{l,j,k} \triangleq 2^{j/2}\psi_l(2^j t - k)$, $j, k \in \mathbf{Z}$, $l = 1, 2, \ldots, N$ form an orthonormal basis for $L^2(\mathbf{R})$. Let $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$ be their corresponding $N \times N$ matrix quadrature mirror filters with $N \times N$ impulse response constant matrices H_k and G_k, $k \in \mathbf{Z}$, respectively. Let

$$\Phi(t) \triangleq (\phi_1(t), \ldots, \phi_N(t))^T, \quad \Psi(t) \triangleq (\psi_1(t), \ldots, \psi_N(t))^T.$$

Then, we have the following matrix dilation equations.

$$\Phi(t) = 2\sum_k H_k \Phi(2t-k) \tag{2.1}$$

$$\Psi(t) = 2\sum_k G_k \Phi(2t-k). \tag{2.2}$$

The orthogonality implies

$$\mathbf{H}(\omega)\mathbf{H}^\dagger(\omega) + \mathbf{H}(\omega+\pi)\mathbf{H}^\dagger(\omega+\pi) = I_N \tag{2.3}$$

$$\mathbf{G}(\omega)\mathbf{G}^\dagger(\omega) + \mathbf{G}(\omega+\pi)\mathbf{G}^\dagger(\omega+\pi) = I_N \tag{2.4}$$

$$\mathbf{H}(\omega)\mathbf{G}^\dagger(\omega) + \mathbf{H}(\omega+\pi)\mathbf{G}^\dagger(\omega+\pi) = 0_N \tag{2.5}$$

where † means the complex conjugate transpose, and I_N and 0_N denote the $N \times N$ identity and the all-zero matrix, respectively.

For each fixed $j \in \mathbf{Z}$, let V_j be the closure of the linear span of $\phi_{l,j,k} \triangleq 2^{j/2}\phi_l(2^j t - k)$, $l = 1, 2, \ldots, N$, $k \in \mathbf{Z}$. Then, the spaces V_j, $j \in \mathbf{Z}$ form an orthogonal multiresolution analysis for $L^2(\mathbf{R})$.

Let $f \in V_J$; then

$$f(t) = \sum_{l=1}^{N}\sum_{k\in\mathbf{Z}} c_{l,J,k}\phi_{l,J,k}(t) \tag{2.6}$$

$$= \sum_{l=1}^{N}\sum_{k\in\mathbf{Z}} c_{l,J_0,k}\phi_{l,J_0,k}(t) + \sum_{l=1}^{N}\sum_{J_0\le j<J}\sum_{k\in\mathbf{Z}} d_{l,j,k}\psi_{l,j,k}(t) \tag{2.7}$$

where $J_0 < J$, and

$$c_{l,j,k} = \int f(t)\phi_{l,j,k}(t)\,dt$$

and

$$d_{l,j,k} = \int f(t)\psi_{l,j,k}(t)\,dt.$$

Let

$$\mathbf{c}_{j,k} \triangleq (c_{1,j,k}, \ldots, c_{N,j,k})^T$$

and

$$\mathbf{d}_{j,k} \triangleq (d_{1,j,k}, \ldots, d_{N,j,k})^T.$$

Then, by the matrix dilations (2.1)–(2.2)

$$\mathbf{c}_{j-1,k} = \sqrt{2}\sum_n H_n \mathbf{c}_{j,2k+n} \tag{2.8}$$

$$\mathbf{d}_{j-1,k} = \sqrt{2}\sum_n G_n \mathbf{c}_{j,2k+n} \tag{2.9}$$

and

$$\mathbf{c}_{j,n} = \sqrt{2}\sum_k (H_k \mathbf{c}_{j-1,2k+n} + G_k \mathbf{d}_{j-1,2k+n}). \tag{2.10}$$

Thus, to determine the multiwavelet transform coefficients $\mathbf{c}_{J_0,k}$ and $\mathbf{d}_{j,k}$ for $J_0 \le j < J$, $k \in \mathbf{Z}$ from f, it is good enough to determine the coefficients $\mathbf{c}_{J,k}$ for $k \in \mathbf{Z}$ from f.

Unlike single wavelets, where $c_{J,k}$ is proportional to the samples $f(k/2^J)$ when J is large enough due to the lowpass property of a single scaling function, the determination of $\mathbf{c}_{J,k}$ for multiwavelet transforms from the samples of $f(t)$ is not trivial. When the multiscaling functions have the interpolating property, the determination was given in [1] and [16]–[17]. Furthermore, a necessary and sufficient condition for the solvability of $\mathbf{c}_{J,k}$ from the samples of f was also given in [1]. The relationship between the samples of f and the coefficients $\mathbf{c}_{J,k}$ automatically provides a prefiltering for the multiwavelet transform computation from the samples of f. For more details, see [1]. Unfortunately, the prefiltering based on this relationship is usually not orthogonal, which seems to limit the gain in the compression applications.

In order to present our new prefilter design method, i.e., a new relationship between the samples of f and $\mathbf{c}_{J,k}$, let us look at the conventional wavelet transform coefficient computation, which is usually referred as the Mallat algorithm.

Let $\phi(t)$ be a single orthogonal scaling function. Then, for any signal $f(t)$, there exists $J > 0$ such that $f(t)$ can be well

approximated by $\phi_{J,k}(t) \triangleq 2^{J/2}\phi(2^J t - k)$, $k \in \mathbf{Z}$, i.e.,

$$f(t) \approx \sum_k c_{J,k}\phi_{J,k}(t) \tag{2.11}$$

where

$$c_{J,k} = \int f(t)\phi_{J,k}(t)\,dt \propto f\left(\frac{k}{2^J}\right). \tag{2.12}$$

The relationship $\propto$ in the above formula is because of the lowpass property of $\phi(t)$, i.e., $\hat{\phi}(0) = 1$, see, for example, [23]–[25] and [35]. The rest of wavelet transform coefficients can be calculated recursively from $c_{J,k}$. The *key point* for the validation of (2.11)–(2.12) is that the scaling function $\phi(t)$ has the lowpass property, and $\phi(t-k)$, $k \in \mathbf{Z}$ are orthogonal.

Motivated from the above observation, we now want to construct a function $\phi(t)$ from the multiscaling functions $\phi_l(t)$, $l = 1, 2, \ldots, N$ such that $\phi(t)$ has the lowpass property, and its translations $\phi(t-k)$, $k \in \mathbf{Z}$ are orthogonal to each other. Notice that such $\phi(t)$ may not be a scaling function because it may not satisfy any dilation equation. As long as $\phi(t)$ has the lowpass property and the orthogonality, the properties (2.11)–(2.12) hold for a signal f.

Let

$$\phi(t) = \sum_{l=1}^{N}\sum_n a_l[n]\phi_l(t-n) \tag{2.13}$$

where $a_l[n]$ are real constants. Then

$$\hat{\phi}(\omega) = \sum_{l=1}^{N} A_l(\omega)\hat{\phi}_l(\omega) \tag{2.14}$$

where

$$A_l(\omega) = \sum_n a_l[n]e^{-jn\omega}. \tag{2.15}$$

The lowpass property implies

$$\hat{\phi}(0) = \sum_{l=1}^{N} A_l(0)\hat{\phi}_l(0) = 1. \tag{2.16}$$

The orthogonality of $\phi(t-n)$, $n \in \mathbf{Z}$ is equivalent to

$$\sum_n |\hat{\phi}(\omega + 2\pi n)|^2 = 1. \tag{2.17}$$

Write out the right-hand side of (2.17) as

$$\begin{aligned}
&\sum_n |\hat{\phi}(\omega+2\pi n)|^2 \\
&\quad = \sum_n \left(\sum_{l_1=1}^{N} A_{l_1}(\omega)\hat{\phi}_{l_1}(\omega+2\pi n)\right) \\
&\qquad \times \left(\sum_{l_2=1}^{N} A^*_{l_2}(\omega)\hat{\phi}^*_{l_2}(\omega+2\pi n)\right) \\
&\quad = \sum_{l_1=1}^{N}\sum_{l_2=1}^{N} A_{l_1}(\omega)A^*_{l_2}(\omega)\sum_n \hat{\phi}_{l_1}(\omega+2\pi n)\hat{\phi}^*_{l_2}(\omega+2\pi n).
\end{aligned}$$

By the orthognality of $\phi_l(t-n)$, $l = 1, 2, \ldots, N$, $n \in \mathbf{Z}$, it is not hard to see that

$$\sum_n \hat{\phi}_{l_1}(\omega+2\pi n)\hat{\phi}^*_{l_2}(\omega+2\pi n) = \delta(l_1 - l_2).$$

Therefore

$$\sum_n |\hat{\phi}(\omega+2\pi n)|^2 = \sum_{l=1}^{N} |A_l(\omega)|^2.$$

This implies that the orthogonality of $\phi(t-n)$, $n \in \mathbf{Z}$ is equivalent to

$$\sum_{l=1}^{N} |A_l(\omega)|^2 = 1. \tag{2.18}$$

In conclusion, we have proved the following lemma.

Lemma 1: A linear combination $\phi(t)$ in (2.13) of multiscaling functions $\phi_l(t)$ and their translations has the lowpass property and the orthogonality of its translations if and only if the properties (2.16) and (2.18) hold.

We now assume $\phi(t)$ in (2.13) satisfies the lowpass property (2.16) and the orthogonality (2.18). For a given signal $f(t)$, by the lowpass property of $\phi(t)$, there exists a $J > 0$ such that (see, for example, [35, Prop. 5.3.2, p. 142])

$$f(t) \approx \sum_n b_n 2^{J/2}\phi(2^J t - n) \tag{2.19}$$

where

$$b_n = \int f(t)2^{J/2}\phi(2^J t - n)\,dt.$$

An estimate of the difference

$$\left| f(t) - \sum_n b_n 2^{J/2}\phi(2^J t - n) \right|$$

is given in the Appendix. Notice that the only condition on $\phi(t)$ for the relationship (2.12) to hold is the lowpass property, i.e., $\hat{\phi}(0) = 1$. Therefore, similar to (2.12), we have

$$b_n \propto f\left(\frac{n}{2^J}\right), \quad \text{for large } J.$$

Without loss of the generality, we may assume $J = 0$ for simplicity. Then

$$f(t) \approx \sum_n b_n\phi(t-n) \quad \text{and} \quad b_n \propto f(n).$$

From (2.13)

$$\begin{aligned}
f(t) &\approx \sum_n b_n\phi(t-n) \\
&= \sum_n b_n \sum_{l=1}^{N}\sum_m a_l[m]\phi_l(t-n-m) \\
&= \sum_{l=1}^{N}\sum_k \left(\sum_m b_{k-m}a_l[m]\right)\phi_l(t-k).
\end{aligned}$$

This implies that

$$\sum_m b_{k-m}a_l[m] \approx c_{l,0,k}, \quad l = 1, 2, \ldots, N, k \in \mathbf{Z} \tag{2.20}$$

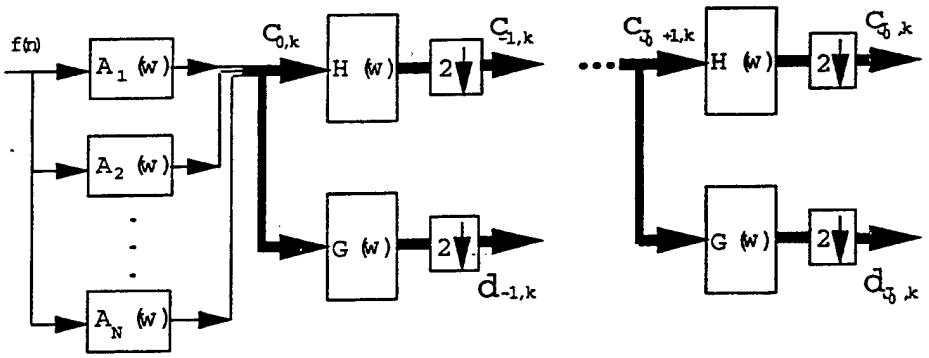

Fig. 1. New prefiltering: Decomposition.

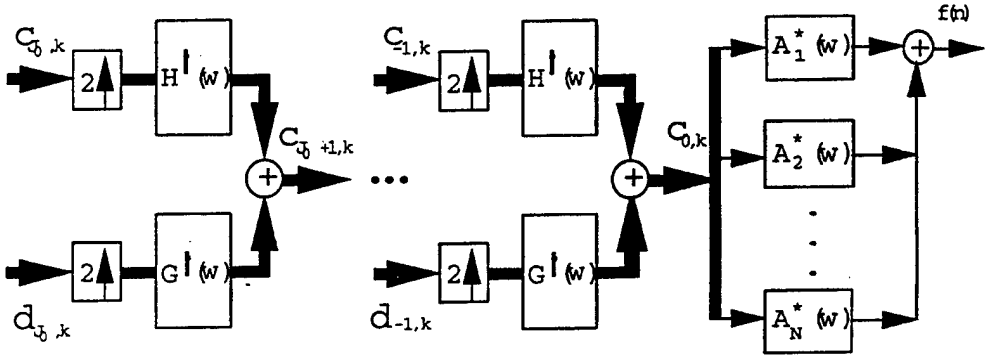

Fig. 2. New prefiltering: Reconstruction.

where $b_n \propto f(n)$, $n \in \mathbf{Z}$. The above result (2.20) suggests the following new relationship, i.e., a new prefilter, between the samples $f(n)$ of $f(t)$ and the multiwavelet transform coefficients $c_{l,0,k}$

$$c_{l,0,k} = \sum_m f(k-m)a_l[m] \tag{2.21}$$

which is shown in Fig. 1.

By the orthogonalities of multiwavelets (2.3)–(2.5) and prefilters (2.18), the reconstruction can be shown in Fig. 2.

The difference between the above prefilter bank and the prefilter bank proposed in [1] is the following. The above prefilter bank is not maximally decimated, i.e., redundancies are introduced. Actually, the number of coefficients in the transform domain is increased by N times. The prefilter bank in [1] is, however, maximally decimated, and no redundancy is introduced. We might want to ask, since we are usually interested in reducing the redundancies, why we need to introduce redundancy here. The answer here is two-fold. First, proper overcomplete (or redundant) transforms plus vector quantizations might perform better than nonredundant transforms. This suggests that including redundancy in the transform might not be a bad idea due to its better tolerance of noise than nonredundant transforms. Second, from our numerical examples, the energy compaction with this new prefiltering is better than the one with Daubechies D_4 wavelet transform after the nonmaximality of the decimation in prefiltering has been taken into account.

Notice that the energy of $f(n)$ is preserved after the whole discrete multiwavelet transform in Fig. 1 is performed due to the orthogonalities of the multiwavelet transform and the prefilter bank, although the prefilter bank is nonmaximally decimated.

Motivated from the above prefiltering and the one in [1], we propose the following general prefiltering for discrete multiwavelet transforms, which is shown in Fig. 3, where $1 \le K \le N$ and the pre/post filterbank shown in Fig. 4 have the perfect reconstruction property. Specifically, when the filterbank in Fig. 4 is paraunitary, the prefiltering in Fig. 3 is orthogonal.

III. Prefilter Design and Examples

In this section, we first study the general N wavelet case and then study the case of $N = 2$. Finally, we look at two examples. One is the Geronimo–Hardin–Massopust 2 wavelet prefilter design, and the other is the prefilter design for one of the 2 wavelets obtained by Chui and Lian in [20].

A. General N Wavelet Prefilter Design

Although, for the general prefilter bank in Fig. 3 (i.e., general K), the interpretation in the previous section does not hold, the design of a prefilter bank $A_l(\omega)$ can be done using the same criterion given in [1], where $K = N$. In the following, we focus on the case of $K = 1$ and use the interpretation in Section II to design the prefilter bank $A_l(\omega)$. Moreover, we are

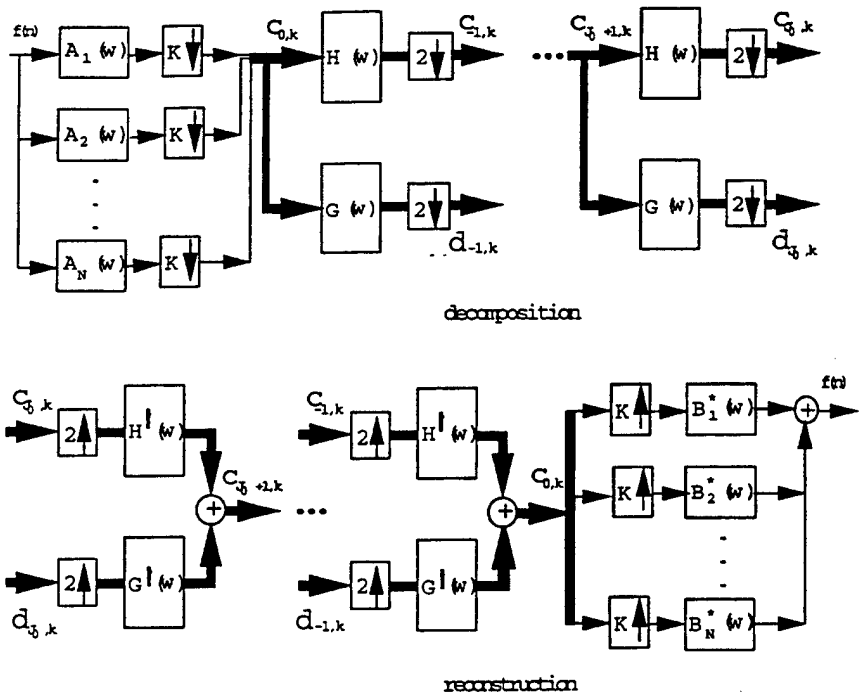

Fig. 3. General prefiltering: Decomposition and reconstruction.

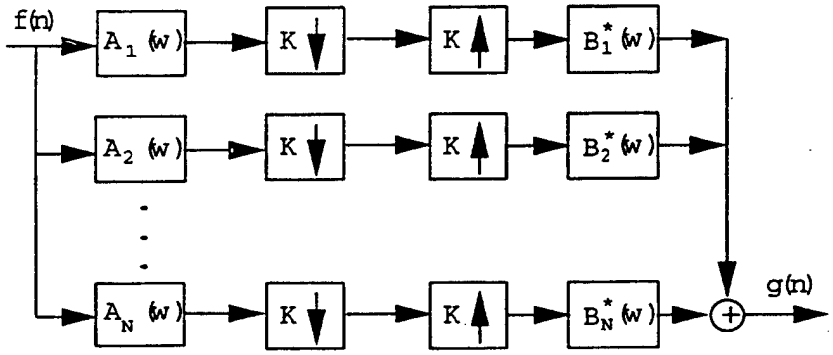

Fig. 4. Pre/post filterbank.

only interested in designing FIR prefilter banks. The lowpass and the orthogonality conditions (2.16) and (2.18) will be used.

Due to its orthogonality, any FIR prefilter bank $A_l(\omega)$ can be factorized as (see, for example, [33], [34])

$$\begin{pmatrix} A_1(\omega) \\ \cdot \\ \cdot \\ A_N(\omega) \end{pmatrix} = U_\rho(\omega)\cdots U_1(\omega)\begin{pmatrix} A_1(0) \\ \cdot \\ \cdot \\ A_N(0) \end{pmatrix} \tag{3.1}$$

where

$$\sum_{l=1}^{N} |A_l(0)|^2 = 1 \tag{3.2}$$

and

$$U_r(\omega) = I_N + (e^{-j\omega} - 1)\mathbf{u}_r^\dagger \mathbf{u}_r \tag{3.3}$$

where $\mathbf{u}_r = (u_{r1}, \ldots, u_{rN})$ and the norm of the vector $\mathbf{u}_r$ is 1, i.e.,

$$\sum_{l=1}^{N} |u_{rl}|^2 = 1.$$

From the matrix dilation equation, we have

$$\begin{pmatrix} \hat{\phi}_1(0) \\ \cdot \\ \cdot \\ \hat{\phi}_N(0) \end{pmatrix} = \mathbf{H}(0)\begin{pmatrix} \hat{\phi}_1(0) \\ \cdot \\ \cdot \\ \hat{\phi}_N(0) \end{pmatrix}. \tag{3.4}$$

When $\mathbf{H}(\omega)$ is known, the vector $(\hat{\phi}(0), \ldots, \hat{\phi}_N(0))$ can be solved. Then, the orthogonality and the lowpass property (2.16) and (2.18) are equivalent to

$$\sum_{l=1}^{N} |A_l(0)|^2 = 1 \quad \text{and} \quad \sum_{l=1}^{N} A_l(0)\hat{\phi}_l(0) = 1. \tag{3.5}$$

The only constraint for the parameters u_{rl} is that they need to be of the unit norm for $r = 1, 2, \ldots, \rho$. The parameter ρ determines the prefilter length and is called the *order* of the prefilter $(A_l(\omega))_{l=1,2,\ldots,N}$. When there is no $U_r(\omega)$ term in (3.1), we set $\rho = 0$, i.e., the order of the prefilter is zero.

Additional conditions may be imposed on the above parameters. An important one is that the combined filters of $A_l(\omega)$ and $\mathbf{H}(\omega)$ need to be lowpass filters, and the combined filters of $A_l(\omega)$ and $\mathbf{G}(\omega)$ need to be highpass filters. The reason for this condition is the same as what was proposed in [1], i.e., we need to keep the "lowpass" part and decompose it again and again but quantize the "highpass" part and therefore keep the "highpass" part as small as possible. This means the "highpass" part needs to be the high-frequency part; otherwise, it will have a lot of energy.

By thinking of the multiscaling vectors as the cascaded version of the "lowpass" vector filter $\mathbf{H}(\omega)$, the new lowpass property (2.16) for the function ϕ means the lowpass property for the combined filters of the prefilters $A_l(\omega)$ and cascaded vector filters $\mathbf{H}(\omega)$. Therefore, the above two lowpass conditions [the new one (2.16) and the old one in [1]] somewhat guarantee the lowpass properties of the all-approximation multiwavelet transform coefficients $\mathbf{c}_{j,k}$ for $J_0 \le j < 0$. The old lowpass condition in [1] is for the lowpass property of the first step decomposition $\mathbf{c}_{-1,k}$ and the new lowpass condition in this paper is for the follow-up decompositions $\mathbf{c}_{j,k}$ for $J_0 \le j < -1$. The old lowpass condition in [1] can be stated as follows.

There are N combined filters of $A_k(\omega)$ and $\mathbf{H}(\omega)$ and N combined filters of $A_l(\omega)$ and $\mathbf{G}(\omega)$. They are

$$H_l(\omega) \triangleq \sum_{k=1}^{N} H_{l,k}(\omega) A_k(\omega), \quad l = 1, 2, \ldots, N \tag{3.6}$$

and

$$G_l(\omega) \triangleq \sum_{k=1}^{N} G_{l,k}(\omega) A_k(\omega), \quad l = 1, 2, \ldots, N \tag{3.7}$$

respectively, where $\mathbf{H}(\omega) = (H_{l,k}(\omega))_{N\times N}$, and $\mathbf{G}(\omega) = (G_{l,k}(\omega))_{N\times N}$. Then, the prefiltering, the first step multiwavelet transform decomposition, and their combined filters can be shown in Fig. 5.

The lowpass property on $H_l(\omega)$ is

$$\sum_{k=1}^{N} H_{l,k}(\pi) A_k(\pi) = 0, \quad l = 1, 2, \ldots, N. \tag{3.8}$$

The highpass property on $G_l(\omega)$ is

$$\sum_{k=1}^{N} G_{l,k}(0) A_k(0) = 0, \quad l = 1, 2, \ldots, N. \tag{3.9}$$

In *conclusion*, the above four conditions [i.e., (3.1), (3.5), (3.8), and (3.9)] need to be imposed on the prefilter design given a multiwavelet.

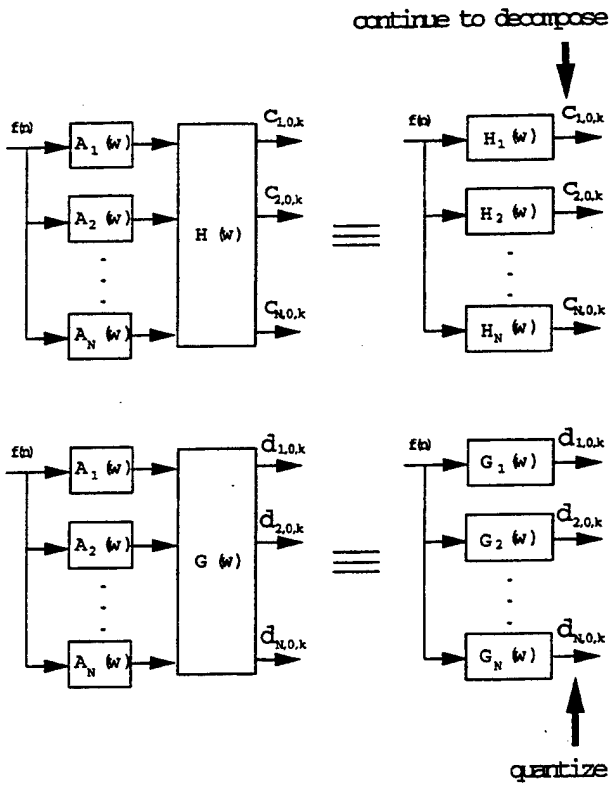

Fig. 5. Combined filters of prefilters and multiwavelet filters.

B. Theory for 2 Wavelets

Since there always exists a solution for (3.4), there exist two real constants a and b such that

$$a\hat{\phi}_1(0) + b\hat{\phi}_2(0) = 0. \tag{3.10}$$

Without loss of generality, we may assume $\hat{\phi}_1(0) = c\hat{\phi}_2(0)$ for a real constant c. Then, by (3.5)

$$cA_1(0) + A_2(0) = 1/\hat{\phi}_2(0) = x, \quad \text{or} \quad A_2(0) = x - cA_1(0) \tag{3.11}$$

and

$$(1 + c^2)A_1^2(0) - 2xcA_1(0) + x^2 - 1 = 0, \quad \text{or}$$
$$A_1(0) = \frac{xc \pm \sqrt{1 + c^2 - x^2}}{1 + c^2} \tag{3.12}$$

where x is an arbitrary constant. This implies that there always exist solutions for (3.5).

When matrix $\mathbf{G}(0)$ has full rank, the only solution for (3.9) is $A_1(0) = A_2(0) = 0$, which does not satisfy (3.5).

When matrix $\mathbf{G}(0)$ does not have full rank, there exist solutions for $A_l(0)$, $l = 1, 2$ in (3.9), i.e., there exist two real constants d and e such that

$$dA_1(0) = eA_2(0). \tag{3.13}$$

Clearly, there exists a solution for $A_1(0)$ and $A_2(0)$ in (3.11)–(3.13).

Now, the only condition left is (3.8). Although the existence of the zeroth-order prefilter $(A_1(\omega), A_2(\omega)) = (A_1(0), A_2(0))$ in (3.8) depends on the form of $(A_1(0), A_2(0))$ and $\mathbf{H}(0)$ (we

will see later that there does not exist any zeroth-order prefilter that satisfies (3.8) for the GHM 2 wavelets, but there does exist for one of the 2 wavelets obtained by Chui and Lian in [20]), we may analyze first order prefilters. In this case

$$\begin{pmatrix} A_1(\omega) \\ A_2(\omega) \end{pmatrix} = \left(I_2 + (e^{-j\omega} - 1) \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} (\cos\theta, \sin\theta) \right) \times \begin{pmatrix} A_1(0) \\ A_2(0) \end{pmatrix} \tag{3.14}$$

and

$$\begin{pmatrix} A_1(\pi) \\ A_2(\pi) \end{pmatrix} = \begin{pmatrix} -\cos 2\theta & -\sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{pmatrix} \begin{pmatrix} A_1(0) \\ A_2(0) \end{pmatrix}$$

where θ is an arbitrary angle.

For the same reason as before, when matrix $\mathbf{H}(\pi)$ has full rank, there are no solutions for (3.5) and (3.8). Therefore, we assume that matrix $\mathbf{H}(\pi)$ does not have full rank. Then, there exist two real constants u, v such that

$$uA_1(\pi) + vA_2(\pi) = 0.$$

By (3.14)

$$(vA_2(0) - uA_1(0))\cos(2\theta) = (uA_2(0) + vA_1(0))\sin(2\theta).$$

Thus, there exists an angle θ such that the above equation holds. This proves the following theorem.

Theorem 1: There exists a first-order prefilter $(A_1(\omega), A_2(\omega))$ that satisfies all conditions [i.e., (3.1), (3.5), (3.8), and (3.9)] if and only if none of matrices $\mathbf{H}(\pi)$ and $\mathbf{G}(0)$ has full rank.

As pointed out by one of the referees of this manuscript, the condition in the above theorem always holds if a constant can be expressed by a linear combination of the translates $\phi_1(t-k)$ and $\phi_2(t-k)$ of two scaling functions $\phi_1(t)$ and $\phi_2(t)$.

C. Design Examples for the GHM 2 Wavelets

We first want to see the Geronimo–Hardin–Massopust 2 wavelets with the following matrix impulse responses of the vector filters $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$, respectively.

$$H_0 = \begin{pmatrix} 3/10 & 2\sqrt{2}/5 \\ -\sqrt{2}/40 & -3/20 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 3/10 & 0 \\ 9\sqrt{2}/40 & 1/2 \end{pmatrix}$$
$$H_2 = \begin{pmatrix} 0 & 0 \\ 9\sqrt{2}/40 & -3/20 \end{pmatrix}, \quad H_3 = \begin{pmatrix} 0 & 0 \\ -\sqrt{2}/40 & 0 \end{pmatrix}$$

and

$$G_0 = \begin{pmatrix} -\sqrt{2}/40 & -3/20 \\ -1/20 & -3\sqrt{2}/20 \end{pmatrix}, \quad G_1 = \begin{pmatrix} 9\sqrt{2}/40 & -1/2 \\ 9/20 & 0 \end{pmatrix}$$
$$G_2 = \begin{pmatrix} 9\sqrt{2}/40 & -3/20 \\ -9/20 & 3\sqrt{2}/20 \end{pmatrix}, \quad G_3 = \begin{pmatrix} -\sqrt{2}/40 & 0 \\ 1/20 & 0 \end{pmatrix}.$$

In this case

$$\mathbf{H}(0) = \begin{pmatrix} \frac{3}{5} & \frac{2\sqrt{2}}{5} \\ \frac{2\sqrt{2}}{5} & \frac{1}{5} \end{pmatrix}, \quad \mathbf{H}(\pi) = \begin{pmatrix} 0 & \frac{2\sqrt{2}}{5} \\ 0 & -\frac{4}{5} \end{pmatrix}$$
$$\mathbf{G}(0) = \begin{pmatrix} \frac{2\sqrt{2}}{5} & -\frac{4}{5} \\ 0 & 0 \end{pmatrix}.$$

From (3.4)

$$\hat{\phi}_1(0) - \sqrt{2}\hat{\phi}_2(0) = 0. \tag{3.15}$$

Solving (3.4) and (3.5), we have

$$A_1(0) = \frac{x\sqrt{2} \pm \sqrt{3 - x^2}}{3} \quad \text{and} \quad A_1(0) = \frac{x \mp \sqrt{3 - x^2}}{3} \tag{3.16}$$

where x is a real constant with $|x| \le \sqrt{3}$. The condition (3.9) implies

$$\sqrt{2}A_2(0) = A_1(0).$$

Therefore, we solve for $A_l(0)$ as

$$A_1(0) = \frac{\sqrt{6}}{3} \quad \text{and} \quad A_2(0) = \frac{\sqrt{3}}{3}.$$

Then, the prefilters in (3.1) can be written as

$$\begin{pmatrix} A_1(\omega) \\ A_2(\omega) \end{pmatrix} = U_\rho(\omega) \cdots U_1(\omega) \begin{pmatrix} \frac{\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} \tag{3.17}$$

where

$$U_r(\omega) = I_2 + (e^{-j\omega} - 1) \begin{pmatrix} u_{r1} \\ u_{r2} \end{pmatrix} (u_{r1}, u_{r2})$$

with $u_{r1}^2 + u_{r2}^2 = 1$ for two real constants u_{r1} and u_{r2}. It is clear that (3.8) implies that the order ρ in (3.17) must be greater than or equal to 1. Since matrices $\mathbf{H}(\pi)$ and $\mathbf{G}(0)$ do not have full rank, by Theorem 1, there exists a first-order prefilter satisfying the conditions. Let us see what it looks like.

$$\begin{pmatrix} A_1(\omega) \\ A_2(\omega) \end{pmatrix} = \left(I_2 + (e^{-j\omega} - 1) \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} (\cos\theta, \sin\theta) \right) \times \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix} \frac{1}{\sqrt{3}} \tag{3.18}$$

where θ is an angle. Thus

$$\begin{pmatrix} A_1(\pi) \\ A_2(\pi) \end{pmatrix} = \begin{pmatrix} -\cos 2\theta & -\sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{pmatrix} \begin{pmatrix} \sqrt{2} \\ 1 \end{pmatrix} \frac{1}{\sqrt{3}}.$$

Therefore, (3.8) implies

$$-\sqrt{2}\sin 2\theta + \cos 2\theta = 0, \quad \text{or} \quad \theta = \frac{1}{2}\arctan\frac{\sqrt{2}}{2}. \tag{3.19}$$

This proves the following theorem.

Theorem 2: The prefilter in (3.18) with the θ in (3.19) satisfies all conditions we want, including the following.

1) the lowpass property of $\phi(t)$, i.e., $\hat{\phi}(0) = 1$;
2) the orthogonality of $\phi(t-n)$, $n \in \mathbf{Z}$ and the orthogonality of the prefilter bank $A_l(\omega)$ for $l = 1, 2$;
3) the lowpass property of the combined filters $H_l(\omega)$ for $l = 1, 2$ of $A_l(\omega)$, $l = 1, 2$, and $\mathbf{H}(\omega)$;
4) the highpass property of the combined filters $G_l(\omega)$ for $l = 1, 2$ of $A_l(\omega)$, $l = 1, 2$, and $\mathbf{G}(\omega)$.

As mentioned earlier, the zeroth-order prefilter, i.e., without any term $U_r(\omega)$ in (3.18), does not satisfy the above property 3), although it satisfies all the rest, i.e., 1), 2), and 4). Notice that the above zeroth-order prefilter was first used in [16] and [17]. When the order ρ of a prefilter increases, better lowpass and highpass combined filters $H_l(\omega)$ and $G_l(\omega)$, respectively, may be expected, and the length of a prefilter also increases. The final version of the two prefilters in (3.18) can be expressed as

$$A_1(\omega) = \frac{\sqrt{2}}{\sqrt{3}}\sin^2\theta - \frac{1}{2\sqrt{3}}\sin 2\theta + \left(\frac{\sqrt{2}}{\sqrt{3}}\cos^2\theta + \frac{1}{2\sqrt{3}}\sin 2\theta\right)e^{-j\omega} \quad (3.20)$$

$$A_2(\omega) = -\frac{1}{\sqrt{6}}\sin 2\theta + \frac{1}{\sqrt{3}}\cos^2\theta + \left(\frac{1}{\sqrt{6}}\sin 2\theta + \frac{1}{\sqrt{3}}\sin^2\theta\right)e^{-j\omega}. \quad (3.21)$$

D. Another Design Example

The second example of 2 wavelets is obtained by Chui and Lian in [20]. The matrix impulse responses are

$$H_0 = \frac{1}{2}\begin{pmatrix} 1/2 & 1/2 \\ -\sqrt{7}/4 & -\sqrt{7}/4 \end{pmatrix}, \quad H_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$
$$H_2 = \frac{1}{2}\begin{pmatrix} 1/2 & -1/2 \\ \sqrt{7}/4 & -\sqrt{7}/4 \end{pmatrix}$$

and

$$G_0 = \frac{1}{2}\begin{pmatrix} -1/2 & -1/2 \\ 1/4 & 1/4 \end{pmatrix}, \quad G_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & \sqrt{7}/2 \end{pmatrix}$$
$$G_2 = \frac{1}{2}\begin{pmatrix} -1/2 & 1/2 \\ -1/4 & 1/4 \end{pmatrix}.$$

The multiscaling and multiwavelet functions are supported in $[0, 2]$ and have symmetry and certain smoothness. It is clear that

$$\mathbf{H}(0) = \frac{1}{2}\begin{pmatrix} 2 & 0 \\ 0 & \frac{1-\sqrt{7}}{2} \end{pmatrix}, \quad \mathbf{H}(\pi) = \frac{1}{2}\begin{pmatrix} 0 & 0 \\ 0 & -\frac{1+\sqrt{7}}{2} \end{pmatrix}$$
$$\mathbf{G}(0) = \frac{1}{2}\begin{pmatrix} 0 & 0 \\ 0 & \frac{1+\sqrt{7}}{2} \end{pmatrix}.$$

Conditions (3.9) and (3.5) imply that $A_1(0) = \pm 1$ and $A_2(0) = 0$. In this case, the zeroth-order $(A_1(\omega), A_2(\omega)) = (A_1(0), A_2(0)) = (1, 0)$ already satisfies (3.8). As pointed out by one of the referees of this manuscript, this result holds not only for the above Chui–Lian multiwavelets but for other multiwavelets as well as long as one of two scaling functions is symmetric and the other of two scaling functions is antisymmetric.

E. Numerical Simulations for the Combined Filters $H_l(\omega)$ and $G_l(\omega)$

In this section, we want to illustrate the combined filters $H_l(\omega)$ and $G_l(\omega)$ for $l = 1, 2$ for the GHM 2 wavelets. Three sets of these combined filters are illustrated: without prefiltering [Fig. 6(a) and (b)]; old zeroth-order orthogonal prefiltering in [1] [Fig. 7(a) and (b)]; new orthogonal prefiltering in Theorem 2 [Fig. 8(a) and (b)].

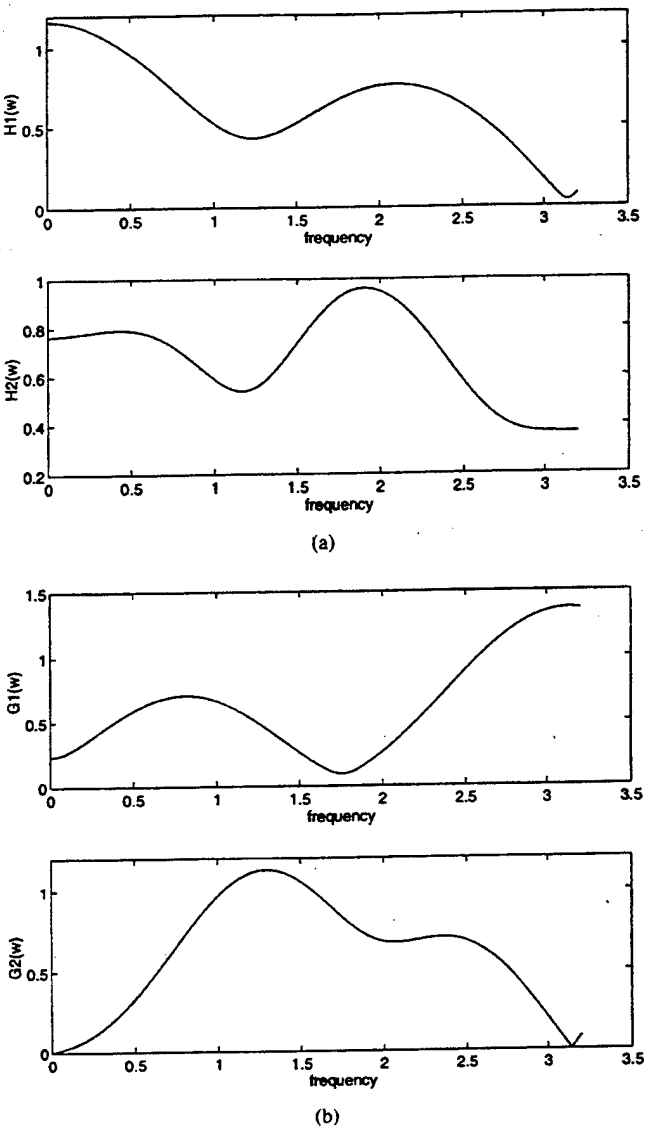

Fig. 6. Combined filters of the GHM 2 wavelets without prefiltering. (a) $|H_l(\omega)|$. (b) $|G_l(\omega)|$.

IV. Numerical Experiments

In this section, we want to see the performance of our new prefiltering scheme through some simple numerical examples. The first test signal is the one hundredth horizontal line of the Cameraman image with size 256×256, which is shown in Fig. 9. Six experiments on energy compaction of the following six transforms are done. The first transform T_1 is the GHM 2 wavelets without prefiltering. The second transform T_2 is the GHM 2 wavelets with the old zeroth-order orthogonal prefiltering with $\epsilon_1 = 1/(10\sqrt{3})$ and $\epsilon_2 = 7/(5\sqrt{6})$ in (3.29) in [1]. The third one T_3 is the Daubechies D_4 wavelets. The forth and the fifth are the GHM 2 wavelets with our new orthogonal prefiltering of the zeroth and the first order, respectively. The sixth transform T_6 is the Chui–Lian multiwavelet transform in Section III-D with the zeroth-order prefiltering $(A_1(\omega), A_2(\omega)) = (1, 0)$. Two step decompositions, i.e., $J_0 =$

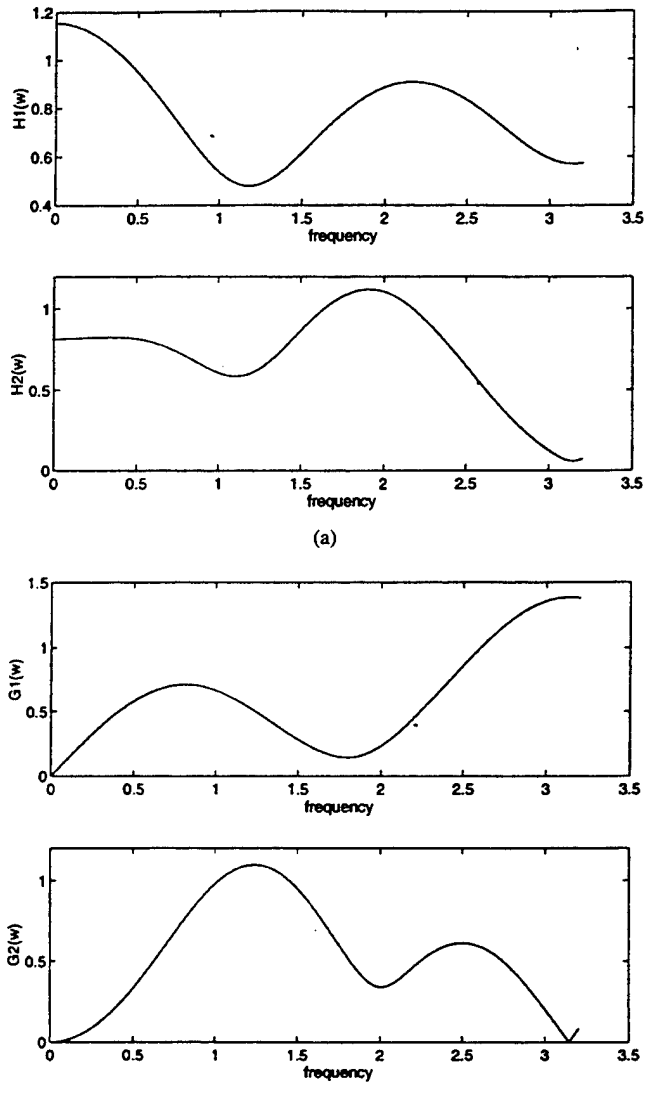

Fig. 7. Combined filters of the GHM 2 wavelets with the old zeroth-order orthogonal prefiltering in [1]. (a) $|H_l(\omega)|$. (b) $|G_l(\omega)|$.

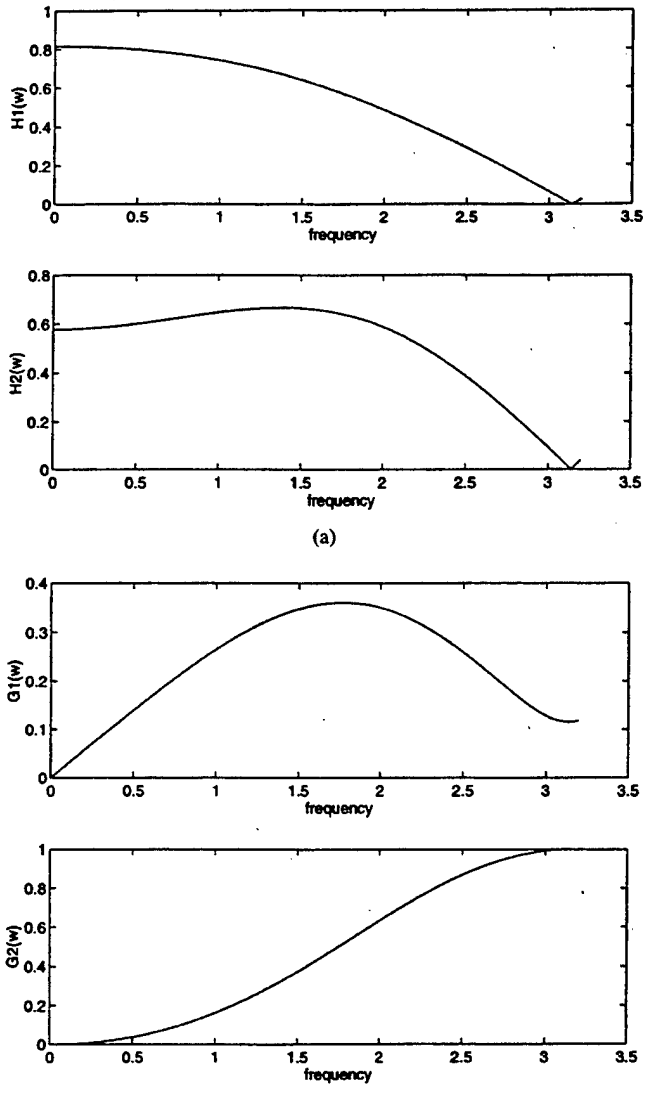

Fig. 8. Combined filters of the GHM 2 wavelets with their first-order orthogonal prefiltering. (a) $|H_l(\omega)|$. (b) $|G_l(\omega)|$.

-2 and $J = 0$, in the first three transforms are performed, where the lowpass part of the transformed signal is of length 64, whereas the bandpass part is of length 192. Since our new prefiltering is nonmaximally decimated and the signal size in the discrete multiwavelet transform domain is twice of the input signal (or the output signals of the first three transforms), three step decompositions, i.e., $J_0 = -3$ and $J = 0$, of the discrete multiwavelet transform with our new prefiltering are performed for the last three transforms, where the length of the lowpass part of the transformed signal is also 64, whereas the length of the bandpass part is $512 - 64 = 448$. Therefore, we have the following energy compaction ratio definitions.

The energy compaction ratioes for the first three transforms T_k for $k = 1, 2, 3$ are defined by

$$r = \frac{\sum_{n=65}^{256} |y[n]|^2}{\sum_{n=1}^{256} |y[n]|^2}$$

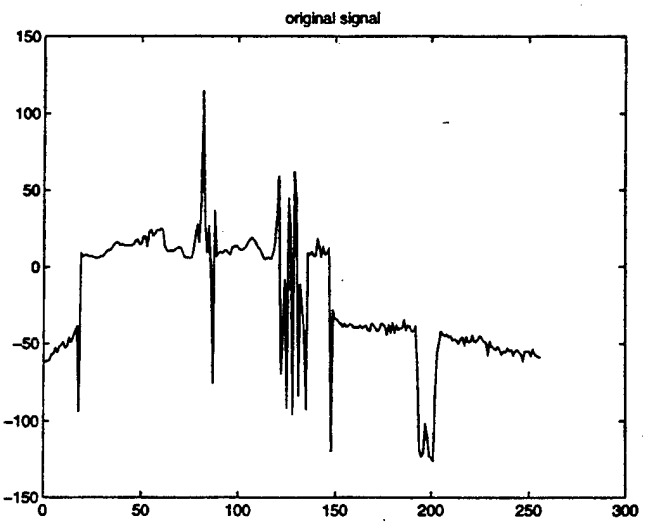

Fig. 9. First test signal.

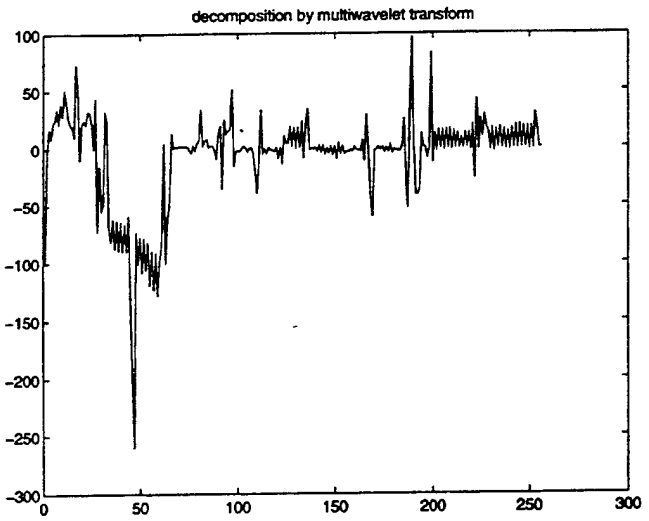

Fig. 10. Decomposition of the first test signal using the GHM 2 wavelets without prefiltering.

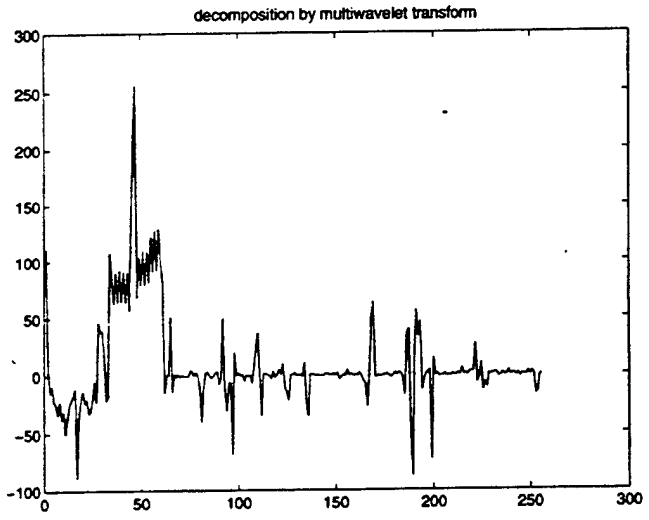

Fig. 11. Decomposition of the first test signal using the GHM 2 wavelets with the old zeroth-order orthogonal prefiltering in [1].

TABLE I
ENERGY COMPACTION RATIO COMPARISON FOR THE FIRST TEST SIGNAL

	τ
GHM 2 wavelets without prefiltering	0.1374
GHM 2 wavelets with the old 0th order orthogonal prefiltering in [1]	0.1247
Daubechies D_4 wavelets	0.1123
GHM 2 wavelets with the new 0th order orthogonal prefiltering	0.0896
GHM 2 wavelets with the new 1th order orthogonal prefiltering	0.0722
Chui-Lian 2 wavelets with the 0th order orthogonal prefiltering	0.0944

where $y[n]$ are the signals in the transform domain. The energy compaction ratioes for the rest three transforms, i.e., with the new prefiltering, are defined by

$$\tau = \frac{\sum_{n=65}^{512} |y[n]|^2}{\sum_{n=1}^{512} |y[n]|^2}.$$

The transformed signals with the first three transforms are shown in Figs. 10–12, respectively. The transformed signals with the new orthogonal prefiltering of the zeroth- order and the first-order for the GHM multiwavelets are shown in Figs. 13 and 14, respectively. The transformed signal with

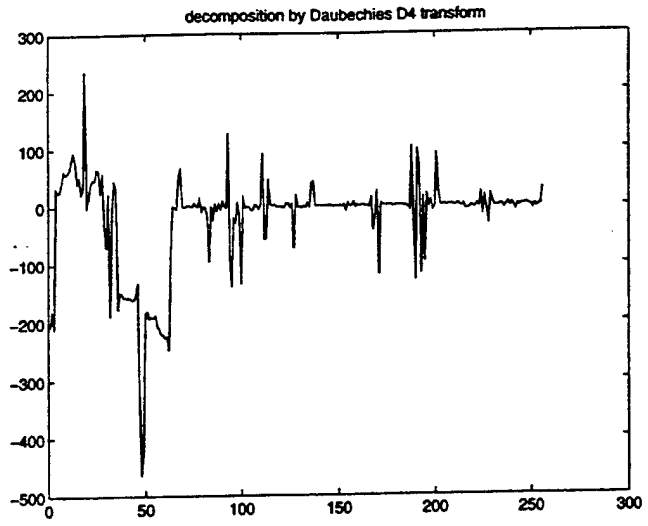

Fig. 12. Decomposition of the first test signal using Daubechies D_4 wavelets.

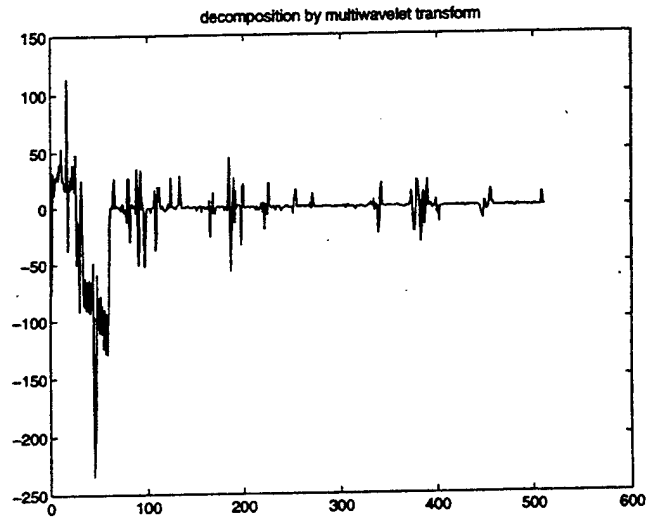

Fig. 13. Decomposition of the first test signal using the GHM 2 wavelets with the new zeroth-order orthogonal prefiltering.

TABLE II
ENERGY COMPACTION RATIO COMPARISON FOR THE SECOND TEST SIGNAL

	τ
Daubechies D_4 wavelets	0.0110
GHM 2 wavelets with the new 1th order orthogonal prefiltering	0.0071
Chui-Lian 2 wavelets with the 0th order orthogonal prefiltering	0.0065

the zeroth-order orthogonal prefiltering for the Chui-Lian multiwavelet is shown in Fig. 15. Their energy compaction ratioes are listed in Table I.

The second test signal is the two hundred and fiftieth horizontal line of the Einstein image with size 256 × 256. The original signal, the transformed signal with transform T_3 (Daubechies D_4 wavelets), the transformed signal with transform T_5 (the GHM 2 wavelets with the first-order orthogonal prefiltering), and transform signal with transform T_6 (the Chui-Lian 2 wavelets with the zeroth-order orthogonal prefiltering) are shown in Figs. 16–19. Their energy compaction ratioes are listed in Table II with the same definitions as above.

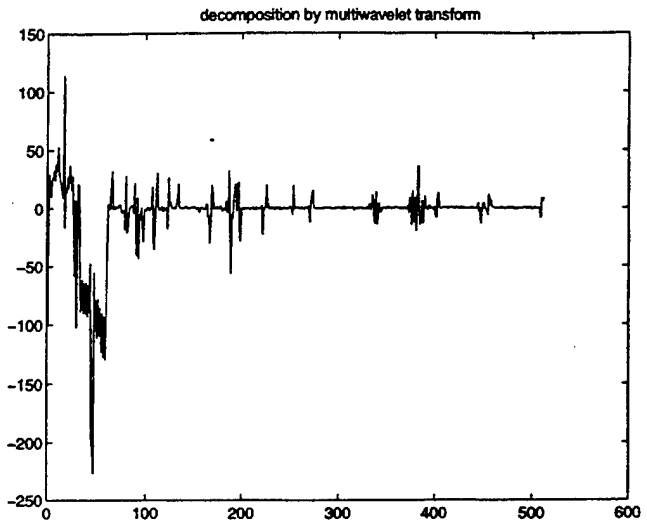

Fig. 14. Decomposition of the first test signal using the GHM 2 wavelets with the new first-order orthogonal prefiltering.

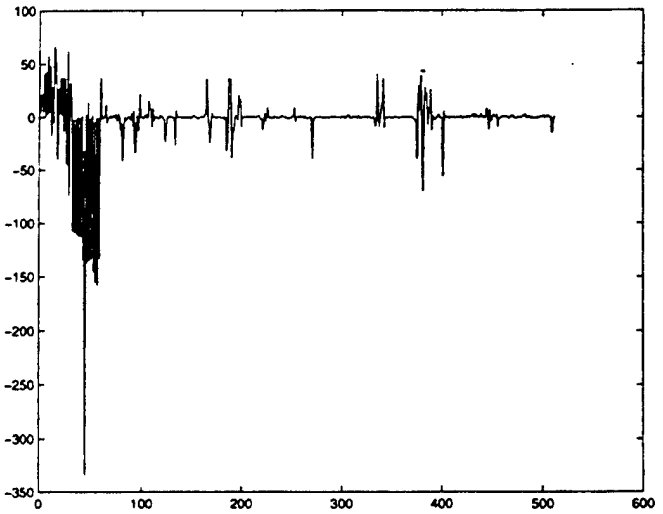

Fig. 15. Decomposition of the first test signal using the Chui-Lian 2 wavelets with the new zeroth-order orthogonal prefiltering.

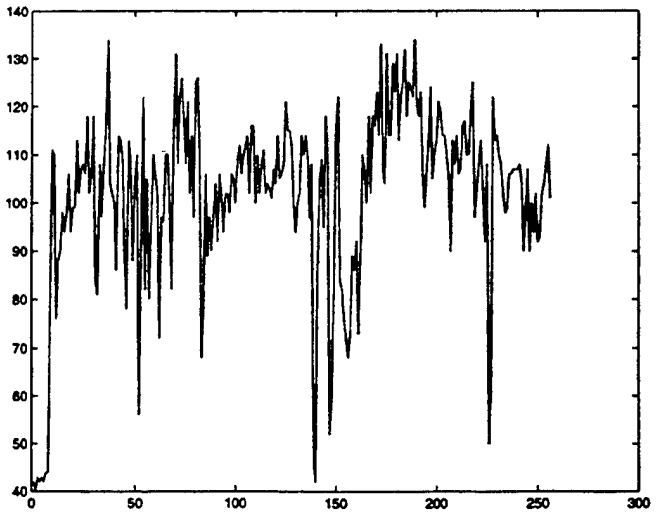

Fig. 16. Second test signal.

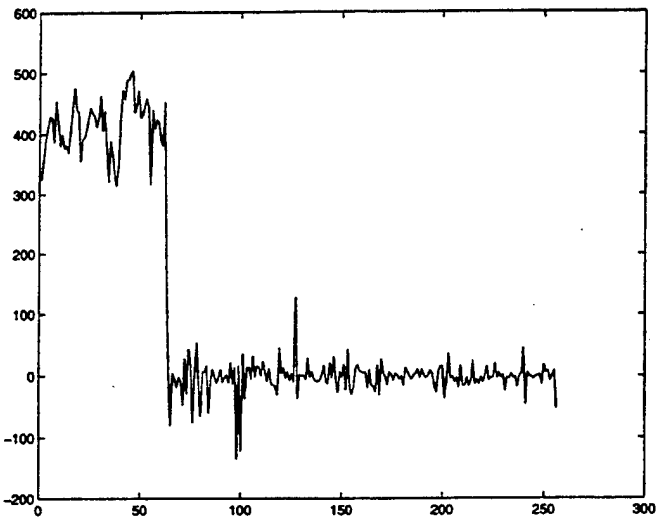

Fig. 17. Decomposition of the second test signal using Daubechies D_4 wavelets.

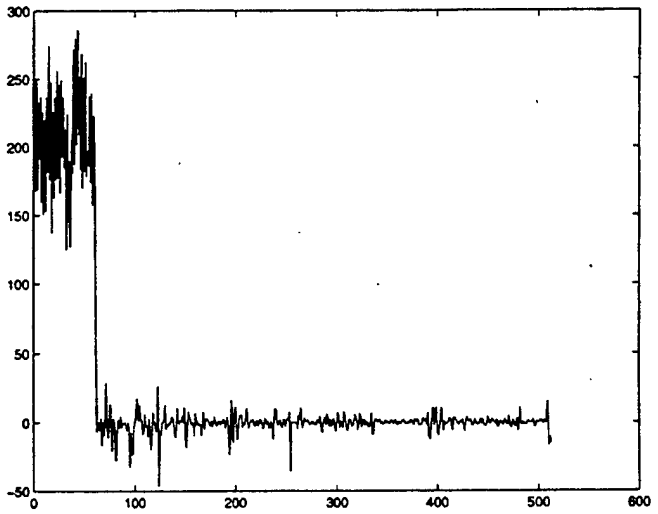

Fig. 18. Decomposition of the second test signal using the GHM 2 wavelets with the new first-order orthogonal prefiltering.

A better energy compaction with the new orthogonal prefilter than with others can be seen from the above tables.

V. Conclusion

In this paper, we have introduced a new prefilter design technique for discrete multiwavelet transforms. The new technique is based on approximating a function with the lowpass property and the orthogonality of their translations by using linear combinations of multiscaling functions and their translations. The new prefiltering is orthogonal but not maximally decimated. It deals with all decomposition steps for discrete multiwavelet transforms, whereas the prefiltering in [1] only focuses on the first step decomposition. The decimation nonmaximality allows one to have more freedom in designing a prefilter so that more desired conditions on the prefilters and the combined filters of the prefilters and multiwavelet vector filters are satisfied. Our numerical examples show that a better energy compaction ratio with the GHM 2 wavelets and the Chui-Lian

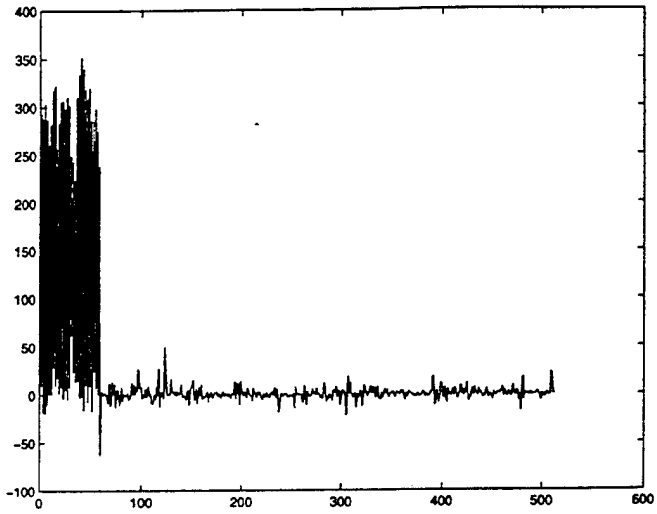

Fig. 19. Decomposition of the second test signal using the Chui-Lian 2 wavelets with the new zeroth-order orthogonal prefiltering.

2 wavelets with the new orthogonal prefiltering than the one with the D_4 wavelet transform is achieved. This suggests the potential applications of discrete multiwavelet transforms in image compression/denoising.

It is known that any nonredundant orthogonal transform keeps the energy. For example, the error energy after the quantization in the transform domain in the compression is equal to the error energy in the reconstruction domain in the decompression. This no longer holds for the redundant prefiltering/postfiltering studied in this paper. In the case when the quantization errors are random, it can be easily shown that the error energy in the reconstruction domain in the decompression is one fourth of the error energy in the transform domain in the compression.

APPENDIX

The error

$$\left| f(t) - \sum_n b_n 2^{J/2} \phi(2^J t - n) \right|$$

where

$$b_n = \int f(t) 2^{J/2} \phi(2^J t - n)\, dt$$

can be estimated as follows. In the Fourier transform domain, the L^2 error can be expressed as

$$\int \left| f(t) - \sum_n b_n 2^{J/2} \phi(2^J t - n) \right|^2 dt = \sqrt{\frac{2^J}{2\pi}} \int \left| \hat{f}(2^J\omega) - \hat{\phi}(\omega) \sum_n \hat{\phi}(-\omega + 2n\pi) \times \hat{f}(2^J(\omega - 2n\pi)) \right|^2 d\omega.$$

When f is bandlimited with bandwidth $2^J\pi$, the error can be simplied as

$$\int \left| f(t) - \sum_n b_n 2^{J/2} \phi(2^J t - n) \right|^2 dt = \frac{1}{2\pi} \int_{-2^J\pi}^{2^J\pi} |\hat{f}(\omega)|^2 \left(1 - \left| \hat{\phi}\left(\frac{\omega}{2^J}\right) \right|^2 \right) d\omega.$$

Notice that $\hat{\phi}(0) = 1$. When J is large enough and the bandwidth W of the signal f is much smaller than $2^J\pi$, i.e., $W \ll 2^J\pi$, then

$$\int \left| f(t) - \sum_n b_n 2^{J/2} \phi(2^J t - n) \right|^2 dt \approx \frac{1}{2\pi} \int_{-W}^{W} |\hat{f}(\omega)|^2 \left(1 - \left| \hat{\phi}\left(\frac{\omega}{2^J}\right) \right|^2 \right) d\omega \approx 0.$$

This is because $\phi(\omega/2^J) \approx 1$ for large J for $|\omega| \le W$.

ACKNOWLEDGMENT

The author would like to thank the referees for their useful comments and suggestions that have improved the clarity of this manuscript.

REFERENCES

[1] X.-G. Xia, J. S. Geronimo, D. P. Hardin, and B. W. Suter, "Design of prefilters for discrete multiwavelet transforms," *IEEE Trans. Signal Processing*, vol. 44, pp. 25–35, Jan. 1996.
[2] J. S. Geronimo, D. P. Hardin, and P. R. Massopust, "Fractal functions and wavelet expansions based on several scaling functions," *J. Approx. Theory*, 1994.
[3] G. Donovan, J. S. Geronimo, D. P. Hardin, and P. R. Massopust, "Construction of orthogonal wavelets using fractal interpolation functions," preprint, 1994.
[4] G. Donovan, J. S. Geronimo, and D. P. Hardin, "Intertwining multiresolution analysis and the construction of piecewise polynomial wavelets," preprint, 1994.
[5] T. N. T. Goodman, S. L. Lee, and W. S. Tang, "Wavelets in wandering subspaces," *Trans. Amer. Math. Soc.*, vol. 338, pp. 639–654, 1993.
[6] T. N. T. Goodman and S. L. Lee, "Wavelets of multiplicity r," *Trans. Amer. Math. Soc.*, vol. 342, pp. 307–324, Mar. 1994.
[7] L. Hervé, "Multi-resolution analysis of multiplicity d: Applications to dyadic interpolation," *Appl. Comput. Harmon. Anal.*, vol. 1, pp. 299–315, 1994.
[8] X.-G. Xia and B. W. Suter, "Vector-valued wavelets and vector filter banks," *IEEE Trans. Signal Processing*, vol. 44, pp. 508–518, Mar. 1996.
[9] ——, "Multirate filter banks with block sampling," *IEEE Trans. Signal Processing*, vol. 44, pp. 484–496, Mar. 1996.
[10] G. Strang and V. Strela, "Orthogonal multiwavelets with vanishing moments," *J. Opt. Eng.*, vol. 33, pp. 2104–2107, 1994.
[11] G. Strang and V. Strela, "Short wavelets and matrix dilation equations," *IEEE Trans. Signal Processing*, vol. 43, pp. 108–115, Jan. 1995.
[12] G. G. Walter, "Orthogonal finite element multiwavelets," preprint, 1994.
[13] M. Vetterli and G. Strang, "Time-varying filter banks and multiwavelets," in *Proc. Sixth Digital Signal Process. Workshop*, Yosemite, CA, Oct. 1994.
[14] C. Heil, G. Strang, and V. Strela, "Approximation by translates of refinable functions," preprint, 1994.
[15] C. Heil and D. Colella, "Matrix refinement equations: Existence and uniqueness," preprint, 1994.
[16] P. N. Heller *et al.*, "Multiwavelet filter banks for data compression," in *Proc. IEEE ISCAS*, 1995.
[17] V. Strela, P. N. Heller, G. Strang, P. Topiwala, and C. Heil, "The application of multiwavelet filter banks to image processing," submitted for publication.
[18] A. Cohen, I. Daubechies, and G. Plonka, "Regularity of refinable function vectors," preprint, 1995.

[19] G. Plonka and V. Strela, "Construction of multi-scaling functions with approximation and symmetry," preprint, 1995.
[20] C. K. Chui and J.-A. Lian, "A study of orthonormal multi-wavelets," CAT Rep. 351, Cent. Approx. Theory, Texas A&M Univ., College Station, Feb. 1995.
[21] O. Rioul and P. Duhamel, "Fast algorithms for discrete and continuous wavelet transforms," *IEEE Trans. Inform. Theory*, vol. 38, pp. 569–586, Mar. 1992.
[22] M. J. Shensa, "The discrete wavelet transform: Wedding the Atrous and Mallat algorithm," *IEEE Trans. Signal Process.*, vol. 40, pp. 2464–2482, Oct. 1992.
[23] X.-G. Xia, "Topics in wavelet transforms," *Ph.D. Dissertation*, Dept. Elect. Eng.–Syst., Univ. Southern Calif., Los Angeles, 1992.
[24] X.-G. Xia and Z. Zhang, "On sampling theorem, wavelets, and wavelet transforms," *IEEE Trans. Signal Processing*, vol. 41, pp. 3524–3535, Dec. 1993.
[25] X.-G. Xia, C.-C. J. Kuo, and Z. Zhang, "Wavelet coefficient computation with optimal prefiltering," *IEEE Trans. Signal Processing*, vol. 42, pp. 2191–2197, Aug. 1994.
[26] P. Rieder, J. Götze, and J. A. Nossek, "Algebraic design of discrete multiwavelet transforms," in *Proc. IEEE ICASSP*, Adelaide, Australia, Apr. 1994.
[27] ——, "Multiwavelet transforms based on several scaling functions," in *Proc. IEEE Int. Symp. Time–Freq. Time-Scale Anal.*, Philadelphia, PA, Oct. 1994.
[28] A. Aldroubi, "Oblique and biorthogonal multi-wavelet bases with fast-filtering algorithms," in *Proc. SPIE*, San Diego, CA, July 1995, vol. 2569, pp. 15–26.
[29] K. N. Johnson and T. Q. Nguyen, "Lattice structure for multifilters derived from complex-valued scalar filterbanks," in *Proc. SPIE*, Denver, CO, Aug. 1996, vol. 2825.
[30] J. T. Miller and C. C. Li, "Adaptive multiwavelet initialization," preprint, 1996.
[31] D. P. Hardin and D. W. Roach, "Multiwavelet prefilters I: Orthogonal prefilters preserving approximation order $p \leq 2$," preprint, 1997.
[32] J. Lebrun and M. Vetterli, "Balanced multiwavelets," in *Proc. ICASSP*, Munich, Germany, May 1997.
[33] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
[34] G. Strang and T. Q. Nguyen, *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge, 1996.
[35] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

Xiang-Gen Xia (M'97) received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, the M.S. degree in mathematics from Nankai University, Tianjin, China, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1983, 1986, and 1992, respectively.

He was a Lecturer at Nankai University from 1986 to 1988, a Teaching Assistant at the University of Cincinnati, Cincinnati, OH, from 1988 to 1990, a Research Assistant at the University of Southern California from 1990 to 1992, and a Research Scientist at the Air Force Institute of Technology, Wright–Patterson AFB, OH, from 1993 to 1994. He was a Senior/Research Staff Member at Hughes Research Laboratories, Malibu, CA, from 1995 to 1996. In September 1996, he joined the Department of Electrical Engineering, University of Delaware, Newark, where he is currently an Assistant Professor. His current research interests include communication systems including equalization and coding, wavelet transform and multirate filterbank theory and applications, time–frequency analysis and synthesis, and numerical analysis and inverse problems in signal/image processing.

Dr. Xia received the National Science Foundation Faculty Early Career Development (CAREER) Program Award in 1997 and the 1998 Office of Naval research (ONR) Young Investigator Program (YIP) Award. He is currently an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He is also a member of the American Mathematical Society.

Contemporary Mathematics
Volume **216**, 1998

Orthonormal Matrix Valued Wavelets and Matrix Karhunen-Loève Expansion

Xiang-Gen Xia

ABSTRACT. In this paper, we study orthonormal matrix valued wavelets for analyzing matrix (vector) valued signals based on matrix multiresolution analysis. We present a simple sufficient condition on the matrix filter $\mathbf{H}(\omega)$ that leads to orthonormal matrix valued wavelets. The sufficient condition is analogous to the one given by Mallat for scalar valued wavelets. The components at each column of matrix valued wavelets form multiwavelets for a scalar valued signal, where the orthonormality induced from the orthonormal matrix valued wavelets is weaker than the one in the current literature on orthonormal multiwavelets. With the new orthonormality, one is able to construct orthonormal matrix valued wavelets similar to the conventional multiresolution analysis based orthonormal wavelets. Moreover, we show that the new orthonormality provides a complete Karhunen-Loève decomposition for matrix valued signals.

1. Introduction

While wavelets and multiwavelets have been extensively studied lately for a scalar-valued signals, see for example [**1**]-[**17**], there are only a few researches, [**1**], on matrix (vector) valued wavelets for matrix (vector) valued signals. In practice, it is however often to encounter matrix (vector) valued signals, such as video images, multi-spectral images and color images. A significant difference between matrix (vector) valued signals and scalar valued signals is that there are correlations for a matrix (vector) valued signal not only in the time domain but also between its components (or the spatial domain) at a fixed time while there is correlation for a scalar valued signal only in the time domain. The aim of the construction of orthonormal matrix valued wavelets is to decorrelate a matrix (vector) valued signal in both the time and the spatial domains. As a side result, the components at each column of orthonormal matrix valued wavelets also form multiwavelets for scalar valued signals. We will see later that the orthonormality for the multiwavelets generated from orthonormal matrix valued wavelets is weaker than the orthonormality in current literature on orthonormal multiwavelets, [**4**]-[**15**]. In [**1**], orthonormal matrix (vector) multiresolution analysis was introduced for the purpose of constructing

1991 *Mathematics Subject Classification.* Primary 41A58, 46E40, 94A12; Secondary 46C50, 45B05, 46N30.

This work was supported in part by an initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

orthonormal matrix valued wavelets. However, the theory in [1] is not complete in the continuous time case in the sense that there is not a simple sufficient condition on the matrix quadrature mirror filter (MQMF) $\mathbf{H}(\omega)$ that leads to orthonormal matrix valued wavelets.

In this paper, we first re-introduce matrix valued signal spaces and matrix valued multiresolution analysis studied in [1]. We then present a simple sufficient condition on the MQMF $\mathbf{H}(\omega)$ for constructing orthonormal matrix valued wavelets, which basically proves the conjecture proposed in [1]. A connection between orthonormal matrix valued wavelets and orthonormal multiwavelets in the current literature is studied. It can be seen that the orthonormality for the multiwavelets induced from the orthonormality of orthonormal matrix valued wavelets is weaker than the orthonormality for multiwavelets in the recent literature in the continuous time waveform case, see for example [4]-[15], while they are the same in the discrete time filterbank case. The weaker orthonormality in the continuous time case provides a weaker sufficient conditon for constructing multiwavelets with this weaker orthonormality.

In the second part of this paper, we show that the orthonormality studied in this paper for matrix valued signals gives a complete Karhunen-Loève decomposition for matrix valued signals, i.e., this orthonormality provides a complete decorrelation for a matrix valued signals in both the time and the spatial domains.

2. Matrix Valued Signal Space and Multiresolution Analysis

For convenience, we only study $N \times N$ matrix valued signals and wavelets. We introduce some notations first.

2.1. Matrix Valued Signal Space.

Let

$$\mathbf{C}^{N\times N} = \{A:\ A \text{ is an } N \times N \text{ matrix with entries in the complex plane } \mathbf{C}\},$$

and

$$L^2(a,b;\mathbf{C}^{N\times N}) \triangleq \left\{\mathbf{f}(t) = (f_{k,l}(t))_{N\times N}:\ f_{k,l}(t) \in L^2(a,b), 1 \le k,l \le N\right\}.$$

The signal space $L^2(a,b;\mathbf{C}^{N\times N})$ is called a matrix valued signal space. When $a = -\infty$ and $b = \infty$, $L^2(a,b;\mathbf{C}^{N\times N})$ is also denoted by $L^2(\mathbf{R},\mathbf{C}^{N\times N})$.

For any $A \in \mathbf{C}^{N\times N}$ and $\mathbf{f} \in L^2(a,b;\mathbf{C}^{N\times N})$, the products

$$A\mathbf{f},\ \mathbf{f}A \in L^2(a,b;\mathbf{C}^{N\times N}).$$

This implies that the matrix valued signal space $L^2(a,b;\mathbf{C}^{N\times N})$ is defined over $\mathbf{C}^{N\times N}$ and not simply over $\mathbf{C}$.

Let $\|\cdot\|_M$ denote a matrix norm on $\mathbf{C}^{N\times N}$. For each $\mathbf{f} \in L^2(a,b;\mathbf{C}^{N\times N})$, $\|\mathbf{f}\|$ denotes the norm of $\mathbf{f}$ associated with the matrix norm $\|\cdot\|_M$ as

$$\|\mathbf{f}\| \triangleq \left(\int_a^b \|\mathbf{f}(t)\|_M^2 dt\right)^{1/2}. \tag{2.1}$$

For $\mathbf{f} \in L^2(a,b;\mathbf{C}^{N\times N})$, its integration $\int \mathbf{f}(t)dt$ is defined by the integration of its components.

For two matrix valued signals $\mathbf{f},\mathbf{g} \in L^2(a,b;\mathbf{C}^{N\times N})$, $\langle \mathbf{f},\mathbf{g}\rangle$ denotes the integration of the matrix product $\mathbf{f}(t)\mathbf{g}^\dagger(t)$:

$$\langle \mathbf{f},\mathbf{g}\rangle \triangleq \int_{\mathbf{R}} \mathbf{f}(t)\mathbf{g}^\dagger(t)dt, \tag{2.2}$$

where † denotes the conjugate transpose. For convenience, we still call the operation $\langle\ ,\ \rangle$ in (2.2) *inner product* although it is not the inner product in the common sense. With the definition (2.2) it is clear that $\langle \mathbf{f}, \mathbf{g}\rangle = \langle \mathbf{g}, \mathbf{f}\rangle^\dagger$.

A sequence $\Phi_k(t) \in L^2(a, b; \mathbf{C}^{N\times N})$, $k \in \mathbf{Z}$, is called an *orthonormal set* in $L^2(a, b; \mathbf{C}^{N\times N})$ if

$$\langle \Phi_k, \Phi_l\rangle = \delta(k-l)I_N, \quad k, l \in \mathbf{Z}, \tag{2.3}$$

where $\delta(k) = 1$ when $k = 0$ and $\delta(k) = 0$ when $k \neq 0$ and I_N is the $N \times N$ identity matrix. A sequence $\Phi_k(t) \in L^2(a, b; \mathbf{C}^{N\times N})$, $k \in \mathbf{Z}$, is called an *orthonormal basis* for $L^2(a, b; \mathbf{C}^{N\times N})$ if it satisfies (2.3), and moreover, for any $\mathbf{f}(t) \in L^2(a, b; \mathbf{C}^{N\times N})$ there exists a sequence of $N \times N$ constant matrices F_k such that

$$\mathbf{f}(t) = \sum_{k\in\mathbf{Z}} F_k\Phi_k(t), \quad \text{for } t \in [a, b], \tag{2.4}$$

where the multiplication $F_k\Phi_k(t)$ for each fixed t is the $N\times N$ matrix multiplication, and the convergence for the infinite summation is in the sense of the norm $\|\cdot\|$ defined by (2.1) for the matrix valued signal space.

2.2. Matrix Valued Multiresolution Analysis. We next define matrix valued multiresolution analysis, which is similar to the conventional multiresolution analysis.

A matrix valued multiresolution analysis (MMRA) of $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$ is a nested sequence of closed subspaces $\mathbf{V}_j$, $j \in \mathbf{Z}$, of $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$ such that

(i). $\mathbf{V}_j \subset \mathbf{V}_{j+1}$, $j \in \mathbf{Z}$,
(ii). $\cup_{j\in\mathbf{Z}}\mathbf{V}_j$ is dense in $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$ and $\cap_{j\in\mathbf{Z}}\mathbf{V}_j = \{\mathbf{0}\}$, where $\mathbf{0}$ is the all zero matrix,
(iii). $\mathbf{f}(t) \in \mathbf{V}_j$ if and only if $\mathbf{f}(2t) \in \mathbf{V}_{j+1}$, $j \in \mathbf{Z}$,
(iv). There is a $\Phi \in \mathbf{V}_0$ such that its translations $\Phi(t-k)$, $k \in \mathbf{Z}$, form an *orthonormal* basis for $\mathbf{V}_0$.

The above definition for an MMRA is notationally similar to the one for the conventional multiresolution analysis (MRA). We call $\Phi(t)$ a *matrix valued scaling function* (or simply scaling function) for the MMRA $\{\mathbf{V}_j\}$. Since $\Phi(t) \in \mathbf{V}_0 \subset \mathbf{V}_1$, there exist constant $N \times N$ matrices H_k, $k \in \mathbf{Z}$, such that,

$$\Phi(t) = 2\sum_k H_k\Phi(2t-k). \tag{2.5}$$

Let

$$\mathbf{H}(\omega) = \sum_k H_k e^{-ik\omega}. \tag{2.6}$$

Then,

$$\hat{\Phi}(\omega) = \mathbf{H}(\frac{\omega}{2})\hat{\Phi}(\frac{\omega}{2}) = \mathbf{H}(\frac{\omega}{2})\mathbf{H}(\frac{\omega}{4})\cdots\hat{\Phi}(0), \tag{2.7}$$

where it is assumed that $\hat{\Phi}(\omega)$ is continuous at $\omega = 0$. This assumption is satisfied when $\mathbf{H}(\omega)$ has only finite terms and $\mathbf{H}(0) = I_N$. In this paper, for convenience we assume $\hat{\Phi}(0) = I_N$, which makes an important difference between matrix-valued

wavelets and multiwavelets from the matrix scaling equation or refinable equation point of view. By this assumption,

$$\hat{\Phi}(\omega) = \mathbf{H}(\frac{\omega}{2})\mathbf{H}(\frac{\omega}{4})\cdots = \prod_{k=1}^{\infty} \mathbf{H}(\frac{\omega}{2^k}). \tag{2.8}$$

The equation (2.7) implies

$$\mathbf{H}(0) = I_N, \quad \text{or} \sum_k H_k = I_N. \tag{2.9}$$

It is not hard to see that the orthonormality of $\Phi(t-k)$, $k \in \mathbf{Z}$, (or the orthonormality of MMRA $\{\mathbf{V}_j\}$) is equivalent to

$$\sum_k \hat{\Phi}(\omega + 2\pi k)\hat{\Phi}^\dagger(\omega + 2\pi k) = 2\pi I_N, \ \forall \omega \in \mathbf{R}. \tag{2.10}$$

In terms of the filter $\mathbf{H}(\omega)$, the above orthonormality implies

$$\mathbf{H}(\omega)\mathbf{H}^\dagger(\omega) + \mathbf{H}(\omega+\pi)\mathbf{H}^\dagger(\omega+\pi) = I_N, \ \forall \omega \in \mathbf{R}. \tag{2.11}$$

The orthonormality (2.10) is in the continuous time domain for continuous-time waveforms while the one (2.11) is in the discrete time domain for discrete-time filterbanks.

Assume we have the above MMRA and $\mathbf{H}(\omega)$. We now want to construct its corresponding matrix valued wavelets that form an orthonormal basis for the whole matrix valued signal space $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$.

Let $\mathbf{G}(\omega)$ satisfy

$$\mathbf{G}(\omega)\mathbf{H}^\dagger(\omega) + \mathbf{G}(\omega+\pi)\mathbf{H}^\dagger(\omega+\pi) = \mathbf{0}, \ \forall \omega \in \mathbf{R}, \tag{2.12}$$

and

$$\mathbf{G}(\omega)\mathbf{G}^\dagger(\omega) + \mathbf{G}(\omega+\pi)\mathbf{G}^\dagger(\omega+\pi) = I_N, \ \forall \omega \in \mathbf{R}. \tag{2.13}$$

Let

$$\hat{\Psi}(\omega) = \mathbf{G}(\frac{\omega}{2})\hat{\Phi}(\frac{\omega}{2}). \tag{2.14}$$

The following result was proved in [1].

THEOREM 2.1. *Let $\Psi(t)$ be the matrix valued function with its Fourier transform defined in (2.14). Then, its translations $\Psi(t-k)$, $k \in \mathbf{Z}$, form an orthonormal basis for $\mathbf{W}_0 \triangleq \mathbf{V}_1 \ominus \mathbf{V}_0$. Thus, $\Psi_{j,k}(t) \triangleq 2^{j/2}\Psi(2^j t - k)$, $j,k \in \mathbf{Z}$, form an orthonormal basis for the matrix valued signal space $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$.*

The matrix filters $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$ in (2.11)-(2.13) are called *matrix quadrature mirror filters* (MQMF). Given $\mathbf{H}(\omega)$, $\mathbf{G}(\omega)$ can be constructed by the following method.

Let $\hat{\mathbf{H}}(\omega) = (\mathbf{H}(\omega), \mathbf{H}(\omega+\pi))^\dagger$ and $\hat{\mathbf{G}}(\omega) = (\mathbf{G}(\omega), \mathbf{G}(\omega+\pi))^\dagger$. Then, the orthogonality (2.11)-(2.13) is equivalent to the paraunitariness of the $2N \times 2N$ matrix $(\hat{\mathbf{H}}(\omega), \hat{\mathbf{G}}(\omega))$. Let $\mathbf{H}_j(\omega)$ and $\mathbf{G}_j(\omega)$ for $j = 0, 1$ be the polyphase components of $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$, respectively: $\mathbf{F}(\omega) = \mathbf{F}_0(2\omega) + e^{-i\omega}\mathbf{F}_1(2\omega)$, where $\mathbf{F}$ is $\mathbf{H}$ or $\mathbf{G}$. Then, the above paraunitariness is equivalent to the paraunitariness of the matrix $(\tilde{\mathbf{H}}(\omega), \tilde{\mathbf{G}}(\omega))$, where $\tilde{\mathbf{F}}(\omega) = (\mathbf{F}_0(\omega), \mathbf{F}_1(\omega))^\dagger$ for $\mathbf{F} = \mathbf{H}$ or $\mathbf{G}$. Thus, the construction of $\mathbf{G}(\omega)$ in (2.11)-(2.13) is equivalent to the completion of a $2N \times 2N$ paraunitary matrix given its first N orthogonal columns $\tilde{\mathbf{H}}(\omega)$. This completion can be obtained by employing the state-space description, see for example [**20**]-[**22**],

where only the orthogonal completion of a constant orthogonal matrix is needed for the corresponding constant realization matrix.

In the next section, we want to construct orthonormal matrix valued scaling functions $\Phi(t)$ from the orthogonal filter $\mathbf{H}(\omega)$ in (2.11).

3. Construction of Matrix Valued Wavelets

It is known that the conventional scaling functions or MRA can be constructed from QMF $H(\omega)$ and necessary and sufficient conditions have been obtained, [18]-[19]. For matrix valued wavelets, we present the following results. We first present a lemma. In what follows, we are only interested in FIR MQMF $\mathbf{H}(\omega)$, i.e., $\mathbf{H}(\omega)$ is a polynomial matrix of $e^{-i\omega}$.

LEMMA 3.1. *Let $\mathbf{H}(\omega)$ satisfy (2.9) and (2.11). If there exist a constant $C > 0$ and an integer K_0 such that for any $\omega \in (-2^K\pi, 2^K\pi)$ and any $K > K_0$,*

$$\|\prod_{l=1}^{K} \mathbf{H}(\frac{\omega}{2^l})\|_M \le C\|\prod_{l=1}^{\infty} \mathbf{H}(\frac{\omega}{2^l})\|_M, \tag{3.1}$$

then, the solution $\Phi(t)$ in the matrix dilation equation (2.5) is a matrix valued scaling function for an MMRA.

Proof. The assumption of the FIR property on $\mathbf{H}(\omega)$ leads to the finiteness of the right hand side of (3.1). To prove Lemma 3.1 we only need to prove the orthonormality of $\Phi(t-k)$, $k \in \mathbf{Z}$. The rest is similar to the conventional MRA theory, see for example [19].

For an integer $K > 0$, let

$$\mu_K(\omega) = \prod_{l=1}^{K} \mathbf{H}(\frac{\omega}{2^l})\chi_{[-2^K\pi, 2^K\pi]}(\omega).$$

Then,

$$\begin{aligned}
&\int_{\mathbf{R}} \mu_K(\omega)\mu_K^\dagger(\omega)e^{-in\omega}d\omega \\
= &\int_{-2^K\pi}^{2^K\pi} \mathbf{H}(\frac{\omega}{2})\cdots\mathbf{H}(\frac{\omega}{2^K})\mathbf{H}^\dagger(\frac{\omega}{2^K})\cdots\mathbf{H}^\dagger(\frac{\omega}{2})e^{-in\omega}d\omega \\
= &\int_{0}^{2^{K+1}\pi} \mathbf{H}(\frac{\omega}{2})\cdots\mathbf{H}(\frac{\omega}{2^K})\mathbf{H}^\dagger(\frac{\omega}{2^K})\cdots\mathbf{H}^\dagger(\frac{\omega}{2})e^{-in\omega}d\omega \\
= &\int_{0}^{2^K\pi} \prod_{l=1}^{K-1} \mathbf{H}(\frac{\omega}{2^l})\left[\mathbf{H}(\frac{\omega}{2^K})\mathbf{H}^\dagger(\frac{\omega}{2^K}) + \mathbf{H}(\frac{\omega}{2^K}+\pi)\mathbf{H}^\dagger(\frac{\omega}{2^K}+\pi)\right] \\
&\cdot\left(\prod_{l=1}^{K-1} \mathbf{H}(\frac{\omega}{2^l})\right)^\dagger e^{-in\omega}d\omega \\
\overset{\text{by } (2.11)}{=} &\int_{\mathbf{R}} \mu_{K-1}(\omega)\mu_{K-1}^\dagger(\omega)e^{-in\omega}d\omega = \cdots = \int_0^{2\pi} e^{in\omega}d\omega I_N = 2\pi\delta(n)I_N.
\end{aligned}$$

It is clear that $\mu_K(\omega)$ converges to $\Phi(\omega)$ pointwisely in (2.8) since $\mathbf{H}(0) = I_N$ and $\mathbf{H}(\omega)$ is a polynomial matrix of $e^{-i\omega}$. By (3.1),

$$\|\mu_K(\omega)\mu_K^\dagger(\omega) - \hat{\Phi}(\omega)\hat{\Phi}^\dagger(\omega)\|_M \le (C+1)\|\hat{\Phi}(\omega)\hat{\Phi}^\dagger(\omega)\|_M, \ \forall\omega \in \mathbf{R}.$$

By the dominated convergence theorem, we have $\|\mu_K\mu_K^\dagger - \hat{\Phi}\hat{\Phi}^\dagger\| \to 0$ as $K \to \infty$. Therefore,

$$\int \Phi(t)\Phi^*(t-n)dt = \frac{1}{2\pi}\int_{\mathbf{R}} \hat{\Phi}(\omega)\hat{\Phi}^\dagger(\omega)e^{-in\omega}d\omega$$

$$= \frac{1}{2\pi}\lim_{K\to\infty}\int_{\mathbf{R}} \mu_K(\omega)\mu_K^\dagger(\omega)e^{-in\omega}d\omega = \delta(n)I_N.$$

This proves the orthonormality of $\Phi(t-k)$, $k \in \mathbf{Z}$. ♣

We next want to present a sufficient condition on $\mathbf{H}(\omega)$ so that (3.1) is satisfied.

LEMMA 3.2. *Let $\mathbf{H}(\omega)$ be a polynomial matrix of $e^{-i\omega}$ and $\mathbf{H}(0) = I_N$. Then, there exist an integer K_0 and a constant $C > 0$ such that*

$$\|\prod_{l=1}^{K}\mathbf{H}(\frac{\omega}{2^l})\|_M \le C\|\prod_{l=1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\|_M,$$

for $\omega \in (-\pi, \pi)$ and $K > K_0$.

Proof. Since $\mathbf{H}(\omega)$ is a polynomial matrix of $e^{-i\omega}$ and $\hat{\Phi}(0) = I_N$, we have

$$\hat{\Phi}(\omega) = \prod_{k=1}^{\infty}\mathbf{H}(\frac{\omega}{2^k}),$$

and

$$\lim_{\omega\to 0}\|\hat{\Phi}(\omega) - I_N\|_M = 0.$$

Thus, there exists an integer $K_0 > 0$ such that, for $k > K_0$ and $|\omega| < \pi/2$,

$$\|\hat{\Phi}(\frac{\omega}{2^k}) - I_N\|_M \le \epsilon,$$

and

$$\|\hat{\Phi}^{-1}(\frac{\omega}{2^k})\|_M \le \frac{1}{\epsilon}, \quad \text{i.e., } \|\left(\prod_{l=k+1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\right)^{-1}\|_M \le \frac{1}{\epsilon},$$

where ϵ is a small positive constant.

Therefore, for $K > K_0$ and $|\omega| < \pi/2$,

$$\|\prod_{l=1}^{K}\mathbf{H}(\frac{\omega}{2^l})\|_M = \|\prod_{l=1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\left(\prod_{l=K+1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\right)^{-1}\|_M \le C\|\prod_{l=1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\|_M,$$

where $C = 1/\epsilon$. ♣

LEMMA 3.3. *Let $\mathbf{H}(\omega)$ be a polynomial matrix of $e^{-i\omega}$ and $\mathbf{H}(0) = I_N$. If*

$$\inf_{|\omega|<\pi/2}|\lambda(\omega)| > 0$$

for any eigenvalue function $\lambda(\omega)$ of the polynomial matrix $\mathbf{H}(\omega)$ of variable $e^{-i\omega}$, then, there exists a constant $C > 0$ such that, for any $\omega \in (-2^K\pi, 2^K\pi)$,

$$\|\prod_{l=1}^{K}\mathbf{H}(\frac{\omega}{2^l})\|_M \le C\|\prod_{l=1}^{\infty}\mathbf{H}(\frac{\omega}{2^l})\|_M.$$

Proof. For $\omega \in (-2^K\pi, 2^K\pi)$, if $k > K$, then $\omega/2^k \in (-\pi/2, \pi/2)$. By the proof of Lemma 3.2, for $\omega \in (-2^K\pi, 2^K\pi)$,

$$\| \left(\prod_{l=K+K_0+1}^{\infty} \mathbf{H}(\frac{\omega}{2^l}) \right)^{-1} \|_M \leq \frac{1}{\epsilon}.$$

Let us consider the case of $l \in \{K+1, K+2, ..., K+K_0\}$. Let $\delta > 0$ such that $\inf_{|\omega|<\pi/2} |\lambda(\omega)| \geq \delta$ for all eigenvalue functions of the polynomial matrix $\mathbf{H}(\omega)$. Then, $(\lambda(\omega))^{-1}$ is an eigenvalue function of the function matrix $(\mathbf{H}(\omega))^{-1}$ of variable $e^{-i\omega}$ for $|\omega| < \pi/2$. Thus, there exists positive constant C_1, which only depends on δ, such that, for $|\omega| < \pi/2$,

$$\|(\mathbf{H}(\omega))^{-1}\|_M \leq C_1.$$

Therefore, for any $\omega \in (-2^K\pi, 2^K\pi)$,

$$\begin{aligned} \| \prod_{l=1}^{K} \mathbf{H}(\frac{\omega}{2^l}) \|_M &= \| \prod_{l=1}^{\infty} \mathbf{H}(\frac{\omega}{2^l}) \left(\prod_{l=K+K_0+1}^{\infty} \mathbf{H}(\frac{\omega}{2^l}) \right)^{-1} \left(\prod_{l=K+1}^{K+K_0} \mathbf{H}(\frac{\omega}{2^l}) \right)^{-1} \|_M \\ &\leq C \| \prod_{l=1}^{\infty} \mathbf{H}(\frac{\omega}{2^l}) \|_M, \end{aligned}$$

where $C = C_1^{K_0}/\epsilon$. ♣

By combining the above three lemmas, we have proved the following result.

THEOREM 3.4. *Let $\mathbf{H}(\omega)$ be a polynomial matrix of $e^{-i\omega}$ and satisfy (2.9) and (2.11). If*

$$\inf_{|\omega|<\pi/2} |\lambda(\omega)| > 0$$

for any eigenvalue function $\lambda(\omega)$ of the polynomial matrix $\mathbf{H}(\omega)$ of variable $e^{-i\omega}$, then, the solution $\Phi(t)$ in the matrix dilation equation (2.5) is a matrix valued scaling function for an MMRA, and therefore $\Psi_{j,k}(t)$, $j,k \in \mathbf{Z}$, form an orthonormal basis for the matrix valued signal space $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$.

Notice that the above sufficient condition is analogous of the one given by Mallat [18]. With the above sufficient condition, it is not hard to construct nontrivial families of orthonormal matrix valued wavelets. The following is an example.

It is not hard to show that, if $\mathbf{H}(\omega) = \frac{1}{2}(I_N + e^{i\omega}\mathbf{E}(2\omega))$ and $\mathbf{E}(\omega)$ is paraunitary, i.e., $\mathbf{E}(\omega)\mathbf{E}^\dagger(\omega) = I_N$, then $\mathbf{G}(\omega) = e^{-i\omega}\mathbf{H}^\dagger(\omega+\pi)$ and $\mathbf{H}(\omega)$ form a pair of MQMF satisfying (2.11)-(2.13). Such property for $\mathbf{H}(\omega)$ is called the sampling property in [1]. Let $\mathbf{E}(\omega) = \mathbf{U}(\omega)\text{diag}(e^{-ik_1\omega}, \cdots, e^{-ik_N\omega})\mathbf{U}^\dagger(\omega)$ for $k_j = 0$ or 1, where $\mathbf{U}(\omega)$ is an arbitrary paraunitary polynomial matrix and $\mathbf{U}(0) = I_N$. Then, it is not hard to see that the above $\mathbf{H}(\omega)$ and $\mathbf{G}(\omega)$ satisfy (2.11)-(2.13) and the sufficient condition in Theorem 3.4.

4. Connection to Multiwavelets

Let $(\Phi(t))_{lk}$, $(\Psi(t))_{lk}$ and $(\mathbf{V}_j)_{lk}$ be the components at the lth column and kth row of $\Phi(t)$, $\Psi(t)$ and $\mathbf{V}_j$, respectively, $l, k = 1, 2, ..., N$ and $j \in \mathbf{Z}$. Then,

$$(\mathbf{V}_j)_{lk} \subset (\mathbf{V}_{j+1})_{lk}, \text{ and } f(t) \in (\mathbf{V}_j)_{lk} \Longleftrightarrow f(2t) \in (\mathbf{V}_{j+1})_{lk},$$

and

$$\cap_{j\in\mathbf{Z}}(\mathbf{V}_j)_{lk} = \{0\}, \text{ and } \cup_{j\in\mathbf{Z}} (\mathbf{V}_j)_{lk} \text{ is dense in } L^2(\mathbf{R}).$$

Moreover, for any $f_{lk} \in (\mathbf{V}_0)_{lk}$, there exist constants $a_{k_1,m,l,k}$ such that

$$f_{lk}(t) = \sum_{k_1 \in \mathbf{Z}} \sum_{m=1}^{N} a_{k_1,m,l,k} (\Phi(t-k_1))_{mk}, \ t \in \mathbf{R}. \tag{4.1}$$

And, for any $f \in L^2(\mathbf{R})$, there exist constants $a_{j,k_1,l,k}$ such that

$$f(t) = \sum_{j,k_1 \in \mathbf{Z}} \sum_{l=1}^{N} a_{j,k_1,l,k} (\Psi_{jk_1}(t))_{lk}, \ t \in \mathbf{R}, \tag{4.2}$$

where k is any integer with $1 \le k \le N$. This implies the following proposition.

THEOREM 4.1. *Let $\Phi(t)$ be a matrix valued scaling function of an MMRA $\{\mathbf{V}_j\}$ and $\Psi(t)$ be its an associated matrix valued wavelet function. Then, for any fixed k, $1 \le k \le N$, the functions $(\Phi(t))_{lk}$, $l = 1, 2, ..., N$, form multiscaling functions and $(\Psi(t))_{lk}$, $l = 1, 2, ..., N$, form multiwavelets. Moreover, for each pair (l, k), the spaces $(\mathbf{V}_j)_{lk}$, $j \in \mathbf{Z}$, form a multiresolution analysis of multiplicity r_k where r_k is the maximum number of linearly independent functions of $(\Phi(t))_{lk}$, $l = 1, 2, ..., N$.*

For more about multiresolution analysis of multiplicity r, see [**2**]-[**3**]. We next want to study the orthonormality of the column multiscaling functions induced from the orthonormality for matrix valued scaling functions, which is

$$\sum_{m=1}^{N} \int (\Phi(t-\tau_1))_{lm} (\Phi^*(t-\tau_2))_{km} dt = \delta(\tau_1 - \tau_2)\delta(l-k). \tag{4.3}$$

Or,

$$\int (\Phi(t-\tau_1))_{lk} (\Phi^*(t-\tau_2))_{kk} dt + \sum_{m=1, m\neq k}^{N} \int (\Phi(t-\tau_1))_{lm} (\Phi^*(t-\tau_2))_{km} dt = \delta(\tau_1 - \tau_2)\delta(l-k). \tag{4.4}$$

Consider the multiscaling functions from the kth column $(\Phi(t))_{lk}(t)$, $1 \le l \le N$, of $\Phi(t)$. The conventional orthogonality studied in the current literature for multiwavelets is

$$\int (\Phi(t-\tau_1))_{l_1 k} (\Phi^*(t-\tau_2))_{l_2 k} dt = \delta(\tau_1 - \tau_2)\delta(l_1 - l_2). \tag{4.5}$$

We call the orthogonality (4.5) *Orthogonality A*, and the orthogonality (4.4) *Orthogonality B*, for multiscaling functions $(\Phi(t))_{lk}(t)$, $1 \le l \le N$. One can see that the second term in the left hand side of (4.4), Orthogonality B, is the flexibility term over (4.5), Orthogonality A.

LEMMA 4.2. *The conventional Orthogonality A for all column vectors of a matrix valued scaling function implies Orthogonality B induced from the orthogonality for matrix valued scaling functions.*

Proof. To prove (4.4), we only need to prove (4.3), which is

$$\sum_{m=1}^{N} \int (\Phi(t-\tau_1))_{lm} (\Phi^*(t-\tau_2))_{km} dt \overset{(4.5)}{=} \sum_{m=1}^{N} \delta(\tau_1-\tau_2)\delta(l-k) = N\delta(\tau_1-\tau_2)\delta(l-k).$$

♣

Comparing Orthogonality A in (4.5) and Orthogonality B in (4.4) or (4.3), one can see that the former requires the orthogonality for each individual component

in a vector while the later only needs the orthogonality for the vector itself. This implies that Orthogonality B is weaker than Orthogonality A. On the other hand, these two orthogonalities imply the same orthogonality (2.11) for the discrete matrix filterbank $\mathbf{H}(\omega)$.

We now consider a subspace of $L^2(\mathbf{R}, \mathbf{C}^{N\times N})$:

$$L^2(\mathbf{R}, \mathbf{C}^N) = \{\mathbf{f} = (f_{k,l}(t))_{N\times N} \in L^2(\mathbf{R}, \mathbf{C}^{N\times N}) : \ f_{k,l}(t) = 0 \text{ for } 2 \le l \le N\},$$

which is isomorphic to the $N \times 1$ vector valued signal space. We may define its corresponding MAR, scaling functions, wavelet functions similarly. In this case, $\Phi(t) = ((\Phi(t))_{kl})_{N\times N}$ with $(\Phi(t))_{kl} = 0$ for $2 \le l \le N$. Clearly, Orthogonality A and Orthogonality B are equivalent in this case. In other words, Orthogonality A only corresponds to Orthogonality B in a proper subspace of the matrix valued signal space.

With Orthogonality A, necessary and sufficient conditions on $\mathbf{H}(\omega)$ that leads to orthogonal multiwavelets have been obtained, see for example [**15**]. Since the stronger Orthogonality A is used, the necessary and sufficient condition on $\mathbf{H}(\omega)$ is not easy to check or use. However, with the weaker Orthogonality B, the condition on $\mathbf{H}(\omega)$ in Theorem 3.4 is much easier to check so that one is able to use it to construct families of nontrivial orthogonal(B) multiwavelets as studied in Section 3. The basic idea doing this is to embed an $N \times 1$ vector into an $N \times N$ matrix and then use the matrix orthogonality. Another way to interpret this idea is that we lift a one dimensional vector into a two dimensional matrix with additional freedoms to play with, which makes the construction easier. One now might want to ask whether this new Orthogonality B is physically meaningful. The answer is *yes* because it provides a complete decorrelation for matrix valued signals as we shall study in the next section.

5. Matrix Karhunen Loève Expansion

In this section, we show that Orthogonality B provides a complete decorrelation for matrix valued random processes.

5.1. Matrix KL Expansion: Definition. Let $\mathbf{X}(t)$, $t \in [a, b]$ with $-\infty < a < b < \infty$, be a matrix valued random process with finite second moments, i.e.,

$$E(\mathbf{X}^\dagger(t)\mathbf{X}(t)) \in \mathbf{C}^{N\times N},$$

and each path $\mathbf{X}(t) \in L^2(\mathbf{R}; \mathbf{C}^{N\times N})$. Let

$$\mathbf{R}(s,t) \triangleq E(\mathbf{X}^\dagger(s)\mathbf{X}(t)), \ \ s,t \in [a,b]. \tag{5.1}$$

If there exist $\Phi_n(t) \in L^2(a,b;\mathbf{C}^{N\times N})$, $\Lambda_n \in \mathbf{C}^{N\times N}$, $n = 1, 2, ...$, such that

$$\int_a^b \Phi_n(s)\mathbf{R}(s,t)ds = \Lambda_n\Phi_n(t), \ \ n = 1, 2, ..., \ t \in [a,b], \tag{5.2}$$

$$\langle \Phi_n, \Phi_m \rangle = \delta(m-n)I_N, \ \ m, n = 1, 2, ..., \tag{5.3}$$

and

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n \rangle \Phi_n(t), \ \ t \in [a,b], \tag{5.4}$$

then, the expansion of $\mathbf{X}(t)$ in (5.4) is called the *matrix Karhunen-Loève expansion* of $\mathbf{X}(t)$. If the matrix Karhunen-Loève (MKL) expansion of $\mathbf{X}(t)$ exists, then $\mathbf{X}(t)$ is decorrelated into a matrix valued random sequence $\mathbf{Y}_n \triangleq \langle \Phi_n, \mathbf{X} \rangle$ as

$$E(\mathbf{Y}_n \mathbf{Y}_m^\dagger) = \delta(n-m)\Lambda_n, \quad m, n = 1, 2, ... \tag{5.5}$$

The random sequence $\mathbf{Y}_n$, $n = 0, 1, 2, ...$, is called the matrix Karhunen-Loève transform of $\mathbf{X}(t)$.

Notice that when $N = 1$, the above MKL expansions/transforms are reduced to the conventional KL expansions/transforms. The object of this section is to study the MKL expansion of $\mathbf{X}(t)$.

Two special cases were studied in [**23**]-[**24**]. In one, the constant matrix Λ_n in (5.2) was replaced by a scalar value and in the other, $\Phi_n(t)$ in (5.2) was replaced by a scalar-valued function. As mentioned in §3.7 in [**24**], only a few cases satisfy these assumptions, and therefore they are not complete. The main reason for not using the product of two matrices at the right hand side in (5.2) is due to the difficulty of handling the noncommutativity of matrix products.

5.2. The Generalized Hilbert-Schmidt and Mercer's Theorems. Without loss of generality, in what follows we assume $a = 0$ and $b = T > 0$. Let $\mathbf{K}(s,t)$, $s, t \in [0, T]$, be a matrix valued function of two variables in $L^2(0, T; \mathbf{C}^{N\times N})$. In other words, for each $s \in [0, T]$, $K(s, \cdot) \in L^2(0, T; \mathbf{C}^{N\times N})$, and for each $t \in [0, T]$, $K(\cdot, t) \in L^2(0, T; \mathbf{C}^{N\times N})$, and

$$\int_0^T \int_0^T \|\mathbf{K}(s,t)\|_M^2 ds dt < \infty. \tag{5.6}$$

If $\mathbf{K}(s,t)$ satisfies the above conditions, then $\mathbf{K}(s,t)$ is called a matrix Fredholm integral operator. It is clear that a matrix Fredholm integral operator $\mathbf{K}(s,t)$ maps $L^2(0, T; \mathbf{C}^{N\times N})$ into itself:

$$(K\mathbf{f})(t) \triangleq \int_0^T \mathbf{f}(s)\mathbf{K}(s,t)ds \in L^2(0, T; \mathbf{C}^{N\times N}).$$

Let $\Phi(t) \in L^2(0, T; \mathbf{C}^{N\times N})$ with $\langle \Phi, \Phi \rangle = I_N$, and $\Lambda \in \mathbf{C}^{N\times N}$. If the following identity holds:

$$\int_0^T \Phi(s)\mathbf{K}(s,t)ds = \Lambda\Phi(t), \quad t \in [0, T], \tag{5.7}$$

then, $\Phi(t)$ and Λ are called *eigen-matrix-functions* and *eigen-matrix-values* of the operator $\mathbf{K}(s,t)$, respectively.

Notice that the property $\langle \Phi, \Phi \rangle = I_N$ is required in the above definitions of eigen-matrix-functions and eigen-matrix-values, which is different from the scalar-valued case. In the scalar-valued case, if $\phi(t)$ is an eigenfunction associated with an eigenvalue λ for a scalar Fredholm integral operator, then $a\phi(t)$ for any constant $a \neq 0$ is also an eigenfunction associated with λ. It is not known, however, whether the following statement is true: If $\Phi(t)$ is an eigen-matrix-function associated with an eigen-matrix-value Λ for a matrix Fredholm integral operator $\mathbf{K}(s,t)$, then $A\Phi(t)$ or $\Phi(t)A$ for an $N\times N$ matrix $A \in \mathbf{C}^{N\times N}$ is also an eigen-matrix-function associated with Λ for the operator $\mathbf{K}(s,t)$. The difficulty is due to the noncommutativity of matrix multiplications.

A matrix Fredholm integral operator $\mathbf{K}(s,t)$ is called *Hermitian* if $\mathbf{K}(s,t) = \mathbf{K}^\dagger(t,s)$ for $s,t \in [0,T]$. If $\mathbf{K}(s,t)$ is Hermitian and Λ is its eigen-matrix-value, then $\Lambda = \Lambda^\dagger$, i.e., Λ is also Hermitian. This is because

$$\langle \Phi, \mathbf{K}\Phi \rangle = \Lambda = (\langle \Phi, \mathbf{K}\Phi \rangle)^\dagger = \Lambda^\dagger.$$

We associate each matrix Fredholm integral operator $\mathbf{K}(s,t)$ on $[0,T] \times [0,T]$ with the following scalar Fredholm integral operator $K(s,t)$ on $[0,NT] \times [0,NT]$:

$$K(s,t) \triangleq K_{k,l}(s-(k-1)T, t-(l-1)T), \tag{5.8}$$

if $(s,t) \in ((k-1)T, kT] \times ((l-1)T, lT]$, $k,l = 1,2,...,N$, where $K_{k,l}(s,t)$ is the component function of $\mathbf{K}(s,t)$ at the kth row and the lth column. The property (5.6) implies the following properties for $K(s,t)$:

$$\int_0^{NT} \int_0^{NT} |K(s,t)|^2 dt ds < \infty, \tag{5.9}$$

and if $\mathbf{K}(s,t)$ is Hermitian then $K(s,t)$ is also Hermitian, i.e., $K(s,t) = K^*(t,s)$, where * means the complex conjugate.

We now have the following generalized Hilbert-Schmidt theorem.

THEOREM 5.1. *Let $\mathbf{K}(s,t)$, $s,t \in [0,T]$, be a Hermitian matrix Fredholm integral operator and $K(s;t)$, $s,t \in [0,NT]$, be its associated scalar Fredholm integral operator. Let $\lambda_1, \lambda_2, ...,$ all be eigenvalues (including multiples) of $K(s,t)$ with $|\lambda_1| \geq |\lambda_2| \geq \cdots$. Then, an $N \times N$ matrix Λ is an eigen-matrix-value of the operator $\mathbf{K}(s,t)$ if and only if*

$$\Lambda = U\, diag(\lambda_{n_1}, \cdots, \lambda_{n_N}) U^\dagger, \tag{5.10}$$

where U is a certain $N \times N$ unitary matrix, and $n_1, ..., n_N$ are positive integers with $n_1 < n_2 < \cdots < n_N$. Moreover, if the operator $K(s,t)$ doesn't have zero eigenvalue, i.e., $|\lambda_n| > 0$, $n = 1,2,...,$ then, the eigen-matrix-functions $\Phi_n(t)$ corresponding to the eigen-matrix-values $\Lambda_n \triangleq diag(\lambda_{(n-1)N+1}, \cdots, \lambda_{nN})$, $n = 1,2,...,$ form an orthonormal basis for the matrix valued signal space $L^2(0,T;\mathbf{C}^{N\times N})$.

Proof: From the definition of an eigen-matrix-value in (5.7), $U^\dagger \Lambda U$ is an eigen-matrix-value of $\mathbf{K}(s,t)$ if Λ is an eigen-matrix-value of $\mathbf{K}(s,t)$ and U is an $N \times N$ unitary matrix. Thus, to prove Λ in (5.10) is an eigen-matrix-value of $\mathbf{K}(s,t)$, we only need to prove the diagonal matrix $\mathrm{diag}(\lambda_{n_1}, \cdots, \lambda_{n_N})$ is an eigen-matrix-value of $\mathbf{K}(s,t)$. In fact, without loss of generality, we only need to prove Λ_n is an eigen-matrix-value of $\mathbf{K}(s,t)$ for any integer $n > 1$.

Let $\phi_n(t)$, $t \in [0,NT]$, be the eigenfunctions of $K(s,t)$ corresponding to λ_n, $n = 1,2,...,$ i.e., $\phi_n(t)$, $n = 1,2,...,$ form an orthonormal set of $L^2(0,NT;\mathbf{C})$, and

$$\int_0^{NT} \phi_n(s) K(s,t) ds = \lambda_n \phi_n(t), \quad t \in [0,NT]. \tag{5.11}$$

Then, equation (5.11) can be rewritten as

$$\int_0^T \sum_{k=0}^{N-1} \phi_n(s+kT) K(s+kT,t) ds = \lambda_n \phi_n(t), \quad t \in [0,NT]. \tag{5.12}$$

Let $\phi_{k,n}(s) \triangleq \phi_n(s+kT)$, $s \in [0,T]$, $k = 0,1,...,N-1$. Then,

$$\int_0^T \sum_{k=0}^{N-1} \phi_{k,n}(s) K(s+kT,t) ds = \lambda_n \phi_{l,n}(t-lT), \tag{5.13}$$

for $t \in (lT,(l+1)T]$, $l = 0,1,...,N-1$. Let

$$\Phi_n(s) \triangleq \begin{pmatrix} \phi_{0,(n-1)N+1}(s) & \phi_{0,(n-1)N+2}(s) & \cdots & \phi_{0,nN}(s) \\ \phi_{1,(n-1)N+1}(s) & \phi_{1,(n-1)N+2}(s) & \cdots & \phi_{1,nN}(s) \\ & & \cdots & \\ \phi_{N-1,(n-1)N+1}(s) & \phi_{N-1,(n-1)N+2}(s) & \cdots & \phi_{N-1,nN}(s) \end{pmatrix}. \tag{5.14}$$

By (5.8), (5.13) can be rewritten as

$$\int_0^T \Phi_n(s) \mathbf{K}(s,t) dt = \Lambda_n \Phi_n(t), \quad n = 1,2,..., \ t \in [0,T]. \tag{5.15}$$

By the orthonormality of $\phi_n(s)$, $t \in [0,NT]$, it is not hard to see that

$$\langle \Phi_m, \Phi_n \rangle = \delta(m-n) I_N, \quad m,n = 1,2,.... \tag{5.16}$$

Therefore, we have proved that Λ_n, $n = 1,2,...,$ are eigen-matrix-values of the operator $\mathbf{K}(s,t)$.

Conversely, let Λ be an eigen-matrix-value of the operator $\mathbf{K}(s,t)$. By the previous discussion we know that Λ is Hermitian. Thus, there exists a unitary matrix U such that $\Lambda = U \mathrm{diag}(\alpha_1, \cdots, \alpha_N) U^\dagger$ with $|\alpha_1| \geq \cdots \geq |\alpha_N|$. By definition (5.7) of an eigen-matrix-value, $\mathrm{diag}(\alpha_1, \cdots, \alpha_N)$ is also an eigen-matrix-value of $\mathbf{K}(s,t)$, i.e., there is $\Phi(t) \in L^2(0,T;\mathbf{C}^{N\times N})$ with $\langle \Phi, \Phi \rangle = I_N$ such that

$$\int_0^T \Phi(s) \mathbf{K}(s,t) ds = \mathrm{diag}(\alpha_1, \cdots, \alpha_N) \Phi(t), \quad t \in [0,T]. \tag{5.17}$$

Assume $\phi_{m,n}(s)$ is the mth row and the nth column component function of $\Phi(s)$. Let $\phi_n(s) = \phi_{m,n}(s-(m-1)T)$ if $s \in ((m-1)T, mT]$ for $m,n = 1,2,...,N$. By (5.8) and (5.17), the function $\phi_n(s)$ is an eigenfunction of the operator $K(s,t)$ with its corresponding eigenvalue α_n, $n = 1,2,...,N$. Thus, $\alpha_k = \lambda_{n_k}$ for some k with $n_1 < n_2 < \cdots < n_N$. This proves (5.10).

When $K(s,t)$ has no zero eigenvalue, by the scalar Hilbert-Schmidt Theorem (see [**25**]), the eigenfunctions $\phi_n(t)$, $n = 1,2,...,$ form an orthonormal basis for $L^2(0,NT;\mathbf{C}^{N\times N})$. Therefore, any $f(t) \in L^2(0,NT;\mathbf{C})$ can be represented as

$$f(t) = \sum_{n=1}^{\infty} \langle f, \phi_n \rangle \phi_n(t), \quad t \in [0,NT]. \tag{5.18}$$

Similarly, (5.18) can be rewritten as

$$\mathbf{f}(t) = \sum_{n=1}^{\infty} \int_0^T \mathbf{f}(s) (\phi_{0,n}(s), \cdots, \phi_{N-1,n}(s))^\dagger ds (\phi_{0,n}(t), \cdots, \phi_{N-1,n}(t)), \ t \in [0,T],$$

for any $N \times 1$ vector-valued $\mathbf{f} \in L^2(0,T;\mathbf{C}^N)$. By regrouping the above summation, we have

$$\mathbf{f}(t) = \sum_{n=1}^{\infty} \int_0^T \mathbf{f}(s) \Phi_n^\dagger(s) \Phi_n(t) ds, \ t \in [0,T], \ \mathbf{f} \in L^2(0,T;\mathbf{C}^N). \tag{5.19}$$

Extending $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^N)$ to $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^{N\times N})$, we have

$$\mathbf{f}(t) = \sum_{n=1}^{\infty} \langle \mathbf{f}, \Phi_n \rangle \Phi_n(t), \quad t \in [0,T], \ \mathbf{f}(t) \in L^2(0,T;\mathbf{C}^{N\times N}). \tag{5.20}$$

This proves that the sequence $\Phi_n(t)$, $n = 1, 2, ...$, forms an orthonormal basis for $L^2(0,T;\mathbf{C}^{N\times N})$. ♣

From the above proof, the eigen-matrix-function $\Phi_n(t)$ in Theorem 5.1 associated with the eigen-matrix-value Λ_n in Theorem 5.1 is formulated by (5.14), for $n = 1, 2, ...$ We next want to generalize Mercer's Theorem. A matrix Fredholm integral operator $\mathbf{K}(s,t)$ is called *positive* if the $N \times N$ matrix $\langle \mathbf{f}, \mathbf{Kf} \rangle$ for any $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^{N\times N})$ is nonnegative definite, i.e., $\mathbf{x}^\dagger \langle \mathbf{f}, \mathbf{Kf} \rangle \mathbf{x} \geq 0$ for any $\mathbf{x} \in \mathbf{C}^N$.

LEMMA 5.2. *A matrix Fredholm integral operator $\mathbf{K}(s,t)$ is positive if and only if its associated scalar Fredholm integral operator $K(s,t)$ is positive.*

Proof: Writing $\langle f, Kf \rangle$ up, similar to the proof of Theorem 5.1, we have

$$\int_0^{NT} \int_0^{NT} f^*(s) K^*(s,t) f(t) ds dt = \int_0^T \int_0^T \mathbf{f}(t) \mathbf{K}^\dagger(s,t) \mathbf{f}^\dagger(s) dt ds, \tag{5.21}$$

where $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^N)$. On the other hand,

$$\mathbf{x}^\dagger \int_0^T \int_0^T \mathbf{f}(t) \mathbf{K}^\dagger(s,t) \mathbf{f}^\dagger(s) dt ds \mathbf{x} = \int_0^T \int_0^T (\mathbf{x}^\dagger \mathbf{f}(t)) \mathbf{K}^\dagger(s,t) (\mathbf{x}^\dagger \mathbf{f}(s))^\dagger dt ds, \tag{5.22}$$

where $\mathbf{x} \in \mathbf{C}^N$ and $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^{N\times N})$. Since

$$L^2(0,T;\mathbf{C}^N) = \{\mathbf{f}(t)\mathbf{x} : \ \mathbf{x} \in \mathbf{C}^N, \ \mathbf{f} \in L^2(0,T;\mathbf{C}^{N\times N})\},$$

the values in (5.21) are nonnegative for all $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^N)$ is equivalent to that the values in (5.22) are nonnegative for all $\mathbf{x} \in \mathbf{C}^N$ and all $\mathbf{f}(t) \in L^2(0,T;\mathbf{C}^{N\times N})$. This proves Lemma 5.2. ♣

we have the following generalized form of Mercer's Theorem.

THEOREM 5.3. *Let $\mathbf{K}(s,t)$ be a Hermitian matrix Fredholm integral operator. If $\mathbf{K}(s,t)$ is positive and its associated scalar Fredholm integral operator $K(s,t)$ is continuous in $[0,NT] \times [0,NT]$, then*

$$\mathbf{K}(s,t) = \sum_{n=1}^{\infty} \Phi_n^\dagger(s) \Lambda_n \Phi_n(t), \quad s,t \in [0,T], \tag{5.23}$$

where $\Phi_n(t)$ and Λ_n are the same as in Theorem 5.1 and the convergence of the infinite summation is uniform.

Proof: By Lemma 5.2, the operator $K(s,t)$ is also positive. By Mercer's theorem for the operator $K(s,t)$ (see [**25**]),

$$K(s,t) = \sum_{n=1}^{\infty} \phi_n^*(s) \phi_n(t) \lambda_n, \quad s,t \in [0,NT],$$

where ϕ_n, λ_n are eigenfunctions and eigenvalues of $K(s,t)$ and the convergence is uniform. Regrouping the above summation and using the same technique in the proof of Theorem 5.1, (5.23) can be proved. ♣

5.3. Matrix KL Expansions for Continuous-Time Matrix Valued Signals. We now come back to the MKL expansions for continuous-time matrix valued signals.

Let $\mathbf{R}(s,t)$ be the correlation matrix function defined by (5.1) of a matrix valued random process $\mathbf{X}(t)$ with $a=0$ and $b=T$. Assume $\mathbf{R}(s,t) \in L^2(0,T;\mathbf{C}^{N\times N})$. Then $\mathbf{R}(s,t)$ is a Hermitian matrix Fredholm integral operator on $L^2(0,T;\mathbf{C}^{N\times N})$; moreover $\mathbf{R}(s,t)$ is positive. Therefore, we can apply the generalized Hilbert-Schmidt Theorem and the generalized Mercer's Theorem.

Let $R(s,t)$ be the associated scalar Fredholm integral operator of the operator $\mathbf{R}(s,t)$, that is defined by (5.8). Let $\phi_n(t)$, λ_n, $n=1,2,...$, all be eigenfunctions and eigenvalues (including multiples) of the operator $R(s,t)$ with

$$\int_0^{NT} \phi_n(s)R(s,t)ds = \lambda_n\phi_n(t),\ t\in[0,NT],\ n=1,2,..., \tag{5.24}$$

and

$$\int_0^{NT} \phi_m(t)\phi_n^*(t)dt = \delta(m-n),\ \ m,n=1,2,..., \tag{5.25}$$

where $|\lambda_1| \geq |\lambda_2| \geq \cdots$. Since the operator $\mathbf{R}(s,t)$ is positive, by Lemma 5.2, the operator $R(s,t)$ is also positive. Thus, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

Let

$$\Lambda_n \triangleq \mathrm{diag}(\lambda_{(n-1)N+1},\cdots,\lambda_{nN}),\ \ n=1,2,..., \tag{5.26}$$

and, for $t\in[0,T]$, $n=1,2,...$, and $\Phi_n(t)$ defined by (5.14). Then, by Theorem 5.1, its proof and (5.25), $\Phi_n(t)$ is an eigen-matrix-function of the operator $\mathbf{R}(s,t)$ corresponding to the eigen-matrix value Λ_n in (5.26) for $n=1,2,....$ This gives the following first condition on signals so that their MKL expansions exist.

THEOREM 5.4. *Let $\mathbf{X}(t)$, $t\in[0,T]$, be a random process with its correlation matrix function $\mathbf{R}(s,t)\in L^2(0,T;\mathbf{C}^{N\times N})$. If $\lambda_n>0$, $n=1,2,...$, then, for each path of $\mathbf{X}(t)$,*

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n\rangle \Phi_n(t),\ \ t\in[0,T], \tag{5.27}$$

i.e., the MKL expansion of $\mathbf{X}(t)$ exists in the sense (5.2)-(5.4).

The second condition is given by the following theorem.

THEOREM 5.5. *Let $\mathbf{X}(t)$, $t\in[0,T]$, be a random process with its correlation matrix function $\mathbf{R}(s,t)\in L^2(0,T;\mathbf{C}^{N\times N})$. If its associated scalar Fredholm integral operator $R(s,t)$ is continuous in $[0,NT]\times[0,NT]$, then the MKL expansion of $\mathbf{X}(t)$ exists:*

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n\rangle \Phi_n(t),\ \ t\in[0,T], \tag{5.28}$$

where the convergence is in the mean square sense.

The proofs of the above two theorems are straightforward by using the results in Section 5.2.

From Theorems 5.4-5.5, it seems that the MKL expansions of $\mathbf{X}(t)$ depend on the definition of the associated scalar Fredholm integral operator $R(s,t)$ of $\mathbf{R}(s,t)$. One might ask, when the existence of the MKL expansion of $\mathbf{X}(t)$ in the sense

of (5.2)-(5.4) is assumed, whether the MKL expansion of $\mathbf{X}(t)$ changes if the way to define $R(s,t)$ in (5.8) changes. The answer is *NO*. In other words, the MKL expansions (5.27) and (5.28) in Theorems 5.4-5.5 are necessary.

THEOREM 5.6. *Let $\mathbf{X}(t)$, $t \in [0,T]$, be a random process with its correlation matrix function $\mathbf{R}(s,t) \in L^2(0,T;\mathbf{C}^{N\times N})$. If the MKL expansion of $\mathbf{X}(t)$ exists in the sense of (5.2)-(5.4), then the MKL expansion of $\mathbf{X}(t)$ can always be written as*

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n \rangle \Phi_n(t), \quad t \in [0,T], \tag{5.29}$$

where $\Phi_n(t)$, $n = 1, 2...$, are defined in (5.14).

Proof: By (5.2)-(5.4), there exist $\Phi_n'(t) \in L^2(0,T;\mathbf{C}^{N\times N})$ and $\Lambda_n' \in \mathbf{C}^{N\times N}$, $n = 1, 2, ...$, such that

$$\int_0^T \Phi_n'(s)\mathbf{R}(s,t)ds = \Lambda_n'\Phi_n'(t), \ n = 1, 2, ..., \ t \in [0,T],$$

$$\langle \Phi_n', \Phi_m' \rangle = \delta(n-m)I_N, \ m, n = 1, 2, ...,$$

and

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n' \rangle \Phi_n'(t), \quad t \in [0,T]. \tag{5.30}$$

Thus, $\Phi_n'(t)$ is an eigen-matrix-function of the operator $\mathbf{R}(s,t)$ corresponding to the eigen-matrix-value Λ_n' for $n = 1, 2, ...$. By Theorem 5.1, there exist unitary matrices U_n such that $\Lambda_n = U_n^\dagger \Lambda_n' U_n$ for $n = 1, 2, ...$, where the order of the eigenvalues λ_n is rearranged if necessary. Moreover, Λ_n is an eigen-matrix-value of $\mathbf{R}(s,t)$ with its eigen-matrix-function $U_n\Phi_n'(t)$, $n = 1, 2, ...$. Then, similar to the proof of Theorem 5.1, one can show that $\Phi_n(t) = U_n\Phi_n'(t)$, $n = 1, 2, ...$. By (5.30),

$$\mathbf{X}(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, U_n^\dagger \Phi_n \rangle U_n^\dagger \Phi_n(t) = \sum_{n=1}^{\infty} \langle \mathbf{X}, \Phi_n \rangle \Phi_n(t).$$

This proves (5.29). ♣

From Thoerems 5.1-5.6, one can clearly see that a matrix valued random process $\mathbf{X}(t)$ is completely decorrelated in the both time and the spatial domains using Orthogonality B.

6. Conclusion

In this paper, we studied orthonormal matrix valued multiresolution analysis and wavelets. A simple sufficient condition on the matrix filter $\mathbf{H}(\omega)$ that leads to orthonormal matrix valued wavelets is presented, which is analogous to the one given by Mallat in [18] for scalar valued wavelets. This sufficient condition enables us to construct families of nontrivial orthonormal matrix valued wavelets. With orthonormal matrix valued wavelets, one is able to construct multiwavelets with a different orthonormality (called Orthogonality B in this paper) from the one people currently use (called Orthogonality A in this paper). It was shown that Orthogonality B is weaker than Orthogonality A. We believe that this weaker orthogonality makes the sufficient condition simple. The main idea behind it is that one dimensional vectors are lifted to two dimensional matrices, and therefore more

freedoms are avaliable. It was also shown that Orthogonality B provides a complete Karhunen-Loève expansion, i.e., a complete decorrelation, for matrix valued signals.

Acknowledgement

The author would like to thank Mr. Quangcai Zhou for providing Lemmas 3.2-3.3 and their proofs. He also would like to thank for the referees' useful comments and suggestions.

References

[1] X.-G. Xia and B. W. Suter, *Vector-valued wavelets and vector filter banks* IEEE Trans. on Signal Processing **44** (1996), 508–518.

[2] T. N. T. Goodman and S. L. Lee, *Wavelets of multiplicity r* Trans. Amer. Math. Soc.**342** (1994), 307-324.

[3] L. Hervé, *Multi-resolution analysis of multiplicity d: applications to dyadic interpolation* Applied and Computational Harmonic Analysis **1** (1994), 299–315.

[4] J. S. Geronimo, D. P. Hardin and P. R. Massopust, *Fractal functions and wavelet expansions based on several scaling functions* J. Approx. Theory **78** (1994), 373–401.

[5] G. Donovan, J. S. Geronimo, and D. P. Hardin, *Intertwining multiresolution analysis and the construction of piecewise polynomial wavelets* Preprint (1994).

[6] G. Strang and V. Strela, *Orthogonal multiwavelets with vanishing moments* J. Optical Eng. **33** (1994), 2104–2107.

[7] G. Strang and V. Strela, *Short wavelets and matrix dilation equations* IEEE Trans. on Signal Processing **43** (1995), 108–115.

[8] M. Vetterli and G. Strang, *Time-varying filter banks and multiwavelets* Sixth Digital Signal Processing Workshop (1994), Yosemite.

[9] C. Heil and D. Colella, *Matrix refinement equations: existence and uniqueness* J. Fourier Anal. Appl. **2** (1996), 363-377.

[10] X.-G. Xia, J. S. Geronimo, D. P. Hardin, and B. W. Suter, *Design of prefilters for discrete multiwavelet transforms* IEEE Trans. on Signal Processing **44** (1996), 25–35.

[11] V. Strela, P. N. Heller, G. Strang, P. Topiwala, C. Heil, *The application of multiwavelet filter banks to image processing* IEEE Trans. on Image Processing, to appear.

[12] W. Lawton, S. L. Lee, and Z. Shen, *An algorithm for matrix extension and wavelet construction* Math. Comp. **65** (1996), 723–737.

[13] J. Z. Wang, *Stability and linear independence associated with scaling vectors* SIAM J. Math. Anal., to appear.

[14] J. Lian, *Orthogonality criteria for multi-scaling functions* Preprint (1996).

[15] G. Plonka, *Necessary and sufficient conditions for orthonormality of scaling vectors* Preprint (1997).

[16] A. Aldroubi, *Oblique and biorthogonal multi-wavelet bases with fast-filtering algorithms* SPIE Proceedings **2569** (1995), San Diego, 15–26.

[17] P. Rieder, J. Götze, J. A. Nossek, *Multiwavelet transforms based on several scaling functions* Proceedings of IEEE Int. Symp. on Time-Freq. and Time-Scale Anal. (1994).

[18] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbf{R})$ Trans. Amer. Math. Soc. **315** (1989), 69–87.

[19] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[20] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[21] Z. Doğanata, P. P. Vaidyanathan and T. Q. Nguyen, *General synthesis procedures for FIR lossless transfer matrices, for perfect-reconstruction multirate filter bank applications* IEEE Trans. on Acoust. Speech and Signal Proc. **36** (1988), 1561–1574.

[22] X.-G. Xia and B. W. Suter, *FIR parauunitary filter banks given several analysis filters: factorization and construction* IEEE Trans. on Signal Processing **44** (1996), 720–723.

[23] E. J. Kelly and W. L. Root, *A representation of vector-valued random processes* Group Rept. 55-21, revised, MIT, Lincoln Laboratory, April 22, 1960.

[24] H. Van Trees, *Detection, Estimation, and Modulation Theory I*, Wiley, 1968.

[25] F. Riesz and B.Sz. Nagy, *Functional Analysis*, New York, Ungar, 1955.

DEPARTMENT OF ELECTRICAL ENGINEERING, UNIVERSITY OF DELAWARE, NEWARK, DE 19716
E-mail address: xxia@ee.udel.edu

Wavelet transform based watermark for digital images

Xiang-Gen Xia, Charles G. Boncelet and Gonzalo R. Arce

Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716

{*xxia, boncelet, arce*} *@ee.udel.edu*

Abstract: In this paper, we introduce a new multiresolution watermarking method for digital images. The method is based on the discrete wavelet transform (DWT). Pseudo-random codes are added to the large coefficients at the high and middle frequency bands of the DWT of an image. It is shown that this method is more robust to proposed methods to some common image distortions, such as the wavelet transform based image compression, image rescaling/stretching and image halftoning. Moreover, the method is hierarchical.

OCIS codes: (100.0100) Image processing;(110.2960) Image analysis

References

1. R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," Proc. ICIP'94, **2**, 86-90 (1994).
2. I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," Proc. ICIP'96, **3**, 243-246 (1996).
3. J. Zhao and E. Koch, "Embedding robust labels into images for copyright protection," Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies, Vienna, Austria, August 21-25, 242-251 (1995).
4. R. B. Wolfgang and E. J. Delp, "A watermark for digital images," Proc. ICIP'96, **3**, 219-222 (1996).
5. I. Pitas, "A method for signature casting on digital images," Proc. ICIP'96, **3**, 215-218 (1996).
6. N. Nikolaidis and I. Pitas, "Copyright protection of images using robust digital signatures," Proceedings of ICASSP'96, Atlanta, Georgia, May, 2168-2171 (1996).
7. M. D. Swanson, B. Zhu, and A. H. Tewfik, "Transparent robust image watermarking," Proc. ICIP'96, **3**, 211-214 (1996).
8. M. Schneider and S.-F. Chang, "A robust content based digital signature for image authentication," Proc. ICIP'96, **3**, 227-230 (1996).
9. S. Mallat, "Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbf{R})$," Trans. Amer. Math. Soc., **315**, 69-87 (1989).
10. I. Daubechies, "Orthonormal bases of compactly supported wavelets," Comm. on Pure and Appl. Math., **41**, 909-996 (1988).
11. O. Rioul and M. Vetterli, "Wavelets and signal processing," IEEE Signal Processing Magazine, 14-38, (1991).
12. I. Daubechies, Ten Lectures on Wavelets, (SIAM, Philadelphia, 1992).
13. P. P. Vaidyanathan, Multirate Systems and Filter Banks, (Prentice Hall, Englewood Cliffs, NJ, 1993).
14. M. Vetterli and J. Kovačević, Wavelets and Subband Coding, (Prentice Hall, Englewood Cliffs, NJ, 1995).
15. G. Strang and T. Q. Nguyen, Wavelets and Filter Banks, (Wellesley-Cambridge Press, Cambridge, 1996).
16. J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," IEEE Trans. on Signal Processing, **41**, 3445-3462 (1993).
17. R. Ulichney, Digital Halftoning, (MIT Press, Massachusetts, 1987).
18. S. Craver, N. Memon, B-L Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications," IBM Research Report (RC 20755), March 1997.

#7038 - $15.00 US

Received October 14, 1998; Revised November 25, 1998

1. Introduction

With the rapid development of the current information technology, electronic publishing, such as the distribution of digitized images/videos, is becoming more and more popular. An important issue for electronic publishing is copyright protection. Watermarking is one of the current copyright protection methods that have recently received considerable attention. See, for example, [1-8, 18]. Basically, "invisible" watermarking for digital images consists of signing an image with a signature or copyright message such that the message is secretly embedded in the image and there is negligible visible difference between the original and the signed images.

There are two common methods of watermarking: the frequency domain and the spatial domain watermarks, for example [1-8, 18]. In this paper, we focus on frequency domain watermarks. Recent frequency domain watermarking methods are based on the discrete cosine transform (DCT), where pseudo-random sequences, such as M-sequences, are added to the DCT coefficients at the middle frequencies as signatures [2-3]. This approach, of course, matches the current image/video compression standards well, such as JPEG, MPEG1-2, etc. It is likely that the wavelet image/video coding, such as embedded zero-tree wavelet (EZW) coding, will be included in the up-coming image/video compression standards, such as JPEG2000 and MPEG4. Therefore, it is important to study watermarking methods in the wavelet transform domain.

In this paper, we propose a wavelet transform based watermarking method by adding pseudo-random codes to the large coefficients at the high and middle frequency bands of the discrete wavelet transform of an image. The basic idea is the same as the spread spectrum watermarking idea proposed by Cox et. al. in [2]. There are, however, three *advantages* to the approach in the wavelet transform domain. The first advantage is that the watermarking method has multiresolution characteristics and is hierarchical. In the case when the received image is not distorted significantly, the cross correlations with the whole size of the image may not be necessary, and therefore much of the computational load can be saved. The second advantage lies in the following argument. It is usually true that the human eyes are not sensitive to the small changes in edges and textures of an image but are very sensitive to the small changes in the smooth parts of an image. With the DWT, the edges and textures are usually well confined to the high frequency subands, such as HH, LH, HL etc. Large coefficients in these bands usually indicate edges in an image. Therefore, adding watermarks to these large coefficients is difficult for the human eyes to perceive. The third advantage is that this approach matches the emerging image/video compression standards. Our numerical results show that the watermarking method we propose is very robust to wavelet transform based image compressions, such as the embedded zero-tree wavelet (EZW) image compression scheme, and as well as to other common image distortions, such as additive noise, rescaling/stretching, and halftoning. The intuitive reason for the advantage of the DWT approach over the DCT approach in rescaling is as follows. The DCT coefficients for the rescaled image are shifted in two directions from the ones for the original image, which degrades the correlation detection for the watermark. Since the DWT are localized not only in the time but also in the frequency domain [9-15], the degradation for the correlation detection in the DWT domain is not as serious as the one in the DCT domain.

Another difference in this paper with the approach proposed by Cox et. al. in [2] is the watermark detection using the correlation measure. The watermark detection method in [2] is to take the inner product (the correlation at the $\tau = 0$ offset) of the watermark and the difference in the DCT domain of the watermarked image and the original image. Even though both the difference and the watermark are normalized, the

#7038 - $15.00 US

Received October 14, 1998; Revised November 25, 1998

inner product may be small if the difference significantly differs from the watermark although there may be a watermark in the image. In this case, it may fail to detect the watermark. In this paper, we propose to take the correlation at all offsets τ of the watermark and the difference in the DWT domain the watermarked image and the original image in different resolutions. The advantage of this new approach is that, although the peak correlation value may not be large, it is much larger than all other correlation values at other offsets if there is a watermark in the image. This ensures the detection of the watermark even though there is a significant distortion in the watermarked image. The correlation detection method in this paper is a relative measure rather than an absolute measure as in [2].

This paper is organized as follows. In Section 2, we briefly review some basics on discrete wavelet transforms (DWT). In Section 3, we propose our new watermarking method based on the DWT. In Section 4, we implement some numerical experiments in terms of several different image distortions, such as, additive noise, rescaling/stretching, image compression with EZW coding and halftoning.

2. Discrete Wavelet Transform (DWT): A Brief Review

The wavelet transform has been extensively studied in the last decade, see for example [9-16]. Many applications, such as compression, detection, and communications, of wavelet transforms have been found. There are many excellent tutorial books and papers on these topics. Here, we introduce the necessary concepts of the DWT for the purposes of this paper. For more details, see [9-15].

The basic idea in the DWT for a one dimensional signal is the following. A signal is split into two parts, usually high frequencies and low frequencies. The edge components of the signal are largely confined to the high frequency part. The low frequency part is split again into two parts of high and low frequencies. This process is continued an arbitrary number of times, which is usually determined by the application at hand. Furthermore, from these DWT coefficients, the original signal can be reconstructed. This reconstruction process is called the inverse DWT (IDWT). The DWT and IDWT can be mathematically stated as follows.

Let

$$H(\omega) = \sum_k h_k e^{-jk\omega}, \text{ and } G(\omega) = \sum_k g_k e^{-jk\omega}.$$

be a lowpass and a highpass filter, respectively, which satisfy a certain condition for reconstruction to be stated later. A signal, $x[n]$ can be decomposed recursively as

$$c_{j-1,k} = \sum_n h_{n-2k} c_{j,n} \tag{1}$$

$$d_{j-1,k} = \sum_n g_{n-2k} c_{j,n} \tag{2}$$

for $j = J+1, J, ..., J_0$ where $c_{J+1,k} = x[k]$, $k \in \mathbf{Z}$, $J+1$ is the high resolution level index, and J_0 is the low resolution level index. The coefficients $c_{J_0,k}, d_{J_0,k}, d_{J_0+1,k}, ..., d_{J,k}$ are called the DWT of signal $x[n]$, where $c_{J_0,k}$ is the lowest resolution part of $x[n]$ and $d_{j,k}$ are the details of $x[n]$ at various bands of frequencies. Furthermore, the signal $x[n]$ can be reconstructed from its DWT coefficients recursively

$$c_{j,n} = \sum_k h_{n-2k} c_{j-1,k} + \sum_k g_{n-2k} d_{j-1,k}. \tag{3}$$

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

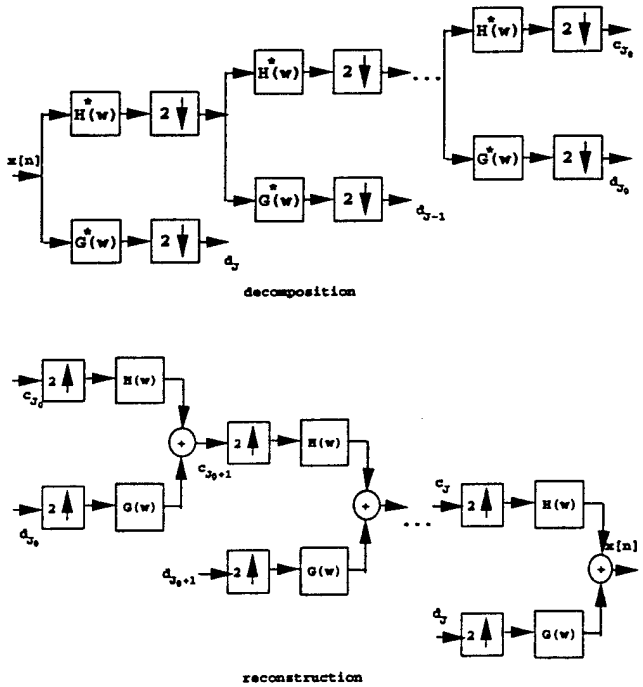

Figure 1. DWT for one dimensional signals.

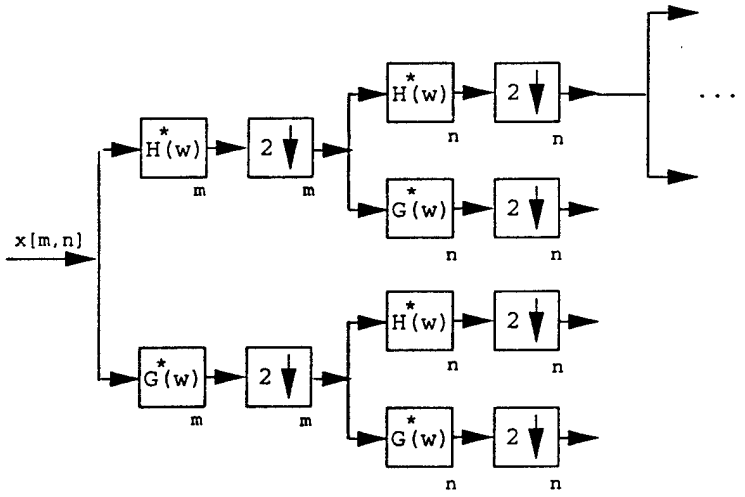

Figure 2. DWT for two dimensional images.

The above reconstruction is called the IDWT of $x[n]$. To ensure the above IDWT and DWT relationship, the following orthogonality condition on the filters $H(\omega)$ and $G(\omega)$ is needed:

$$|H(\omega)|^2 + |G(\omega)|^2 = 1.$$

An example of such $H(\omega)$ and $G(\omega)$ is given by

$$H(\omega) = \frac{1}{2} + \frac{1}{2}e^{-j\omega}, \text{ and } G(\omega) = \frac{1}{2} - \frac{1}{2}e^{-j\omega},$$

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

which are known as the Haar wavelet filters.

The above DWT and IDWT for a one dimensional signal $x[n]$ can be also described in the form of two channel tree-structured filterbanks as shown in Fig. 1. The DWT and IDWT for two dimensional images $x[m,n]$ can be similarly defined by implementing the one dimensional DWT and IDWT for each dimension m and n separately: $DWT_n[DWT_m[x[m,n]]]$, which is shown in Fig. 2. An image can be decomposed into a pyramid structure, shown in Fig. 3, with various band information: such as low-low frequency band, low-high frequency band, high-high frequency band etc. An example of such decomposition with two levels is shown in Fig. 4, where the edges appear in all bands except in the lowest frequency band, i.e., the corner part at the left and top.

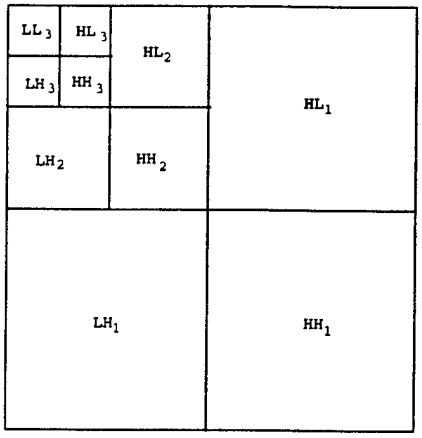

Figure 3. DWT pyramid decomposition of an image.

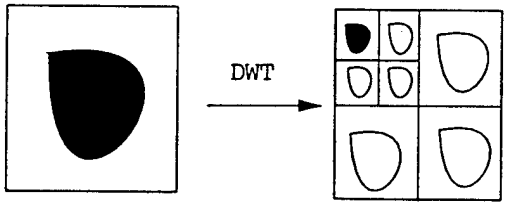

Figure 4. Example of a DWT pyramid decomposition.

3. Watermarking in the DWT Domain

Watermarking in the DWT domain is composed of two parts: encoding and decoding. In the encoding part, we first decompose an image into several bands with a pyramid structure as shown in Figs. 3-4 and then add a pseudo-random sequence (Gaussian noise) to the largest coefficients which are not located in the lowest resolution, i.e., the corner at the left and top, as follows. Let $y[m,n]$ denote the DWT coefficients, which are not located at the lowest frequency band, of an image $x[n,m]$. We add a Gaussian noise sequence $N[m,n]$ with mean 0 and variance 1 to $y[m,n]$:

$$\tilde{y}[m,n] = y[m,n] + \alpha y^2[m,n]N[m,n], \tag{4}$$

where α is a parameter to control the level of the watermark, the square indicates the amplification of the large DWT coeffcients. We do not change the DWT coefficients at the lowest resolution. Then, we take the two dimensional IDWT of the modified DWT

#7038 - $15.00 US

Received October 14, 1998; Revised November 25, 1998

coefficients $\tilde{y}$ and the unchanged DWT coefficients at the lowest resolution. Let $\tilde{x}[m,n]$ denote the IDWT coefficients. For the resultant image to have the same dynamic range as the original image, it is modified as

$$\hat{x}[m,n] = \min(\max(x[m,n]), max\{\tilde{x}[m,n], \min(x[m,n])\}). \quad (5)$$

The operation in (5) is to make the two dimensional data $\tilde{x}[m,n]$ be the same dynamic range as the original image $x[m,n]$. The resultant image $\hat{x}[m,n]$ is the watermarked image of $x[m,n]$. The encoding part is illustrated in Fig. 5(a).

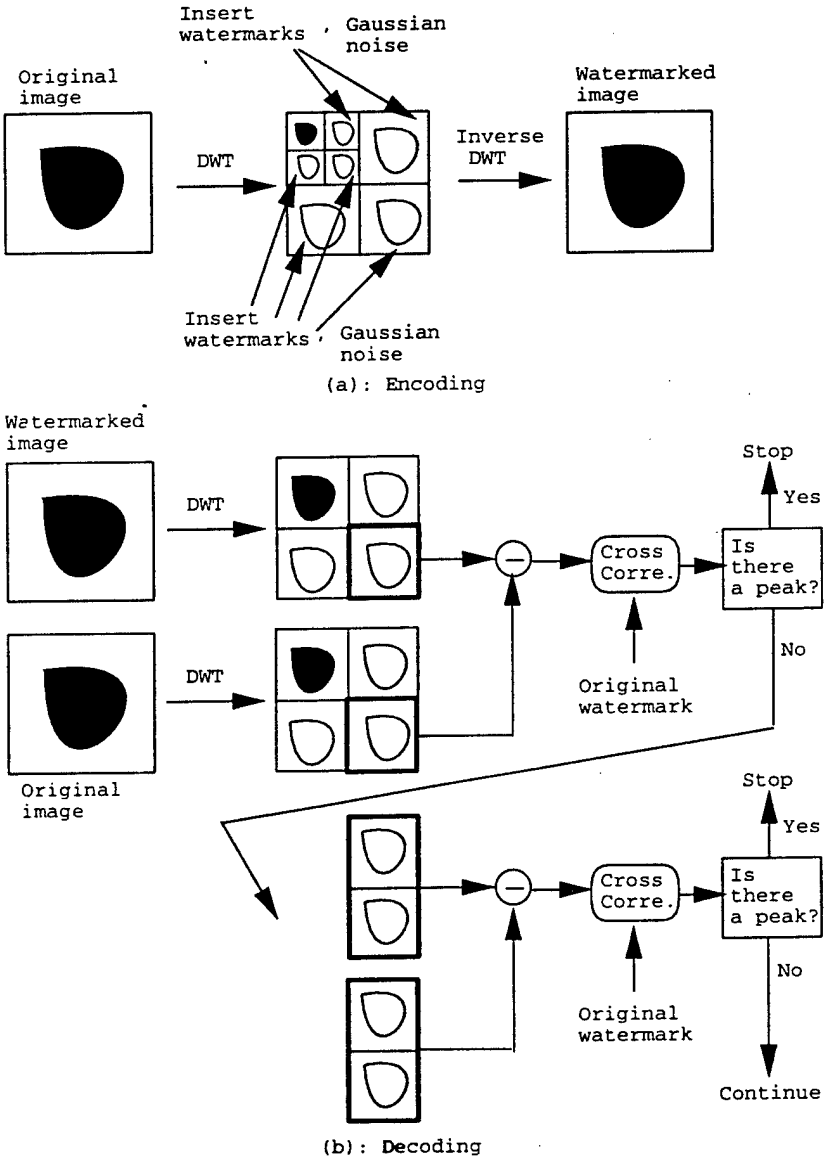

Figure 5. Watermarking in the DWT domain.

The decoding method we propose is hierarchical and described as follows. We first decompose a received image and the original image (it is assumed that the original image is known) with the DWT into four bands, i.e., low-low (LL_1) band, low-high

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

(LH_1) band, high-low (HL_1) band, and high-high (HH_1) band, respectively. We then compare the signature added in the HH_1 band and the difference of the DWT coefficients in HH_1 bands of the received and the original images by calculating their cross correlations. If there is a peak in the cross correlations, the signature is called detected. Otherwise, compare the signature added in the HH_1 and LH_1 bands with the difference of the DWT coefficients in the HH_1 and LH_1 bands, respectively. If there is a peak, the signature is detected. Otherwise, we consider the signature added in the HL_1, LH_1, and HH_1 bands. If there is still no peak in the cross correlations, we continue to decompose the original and the received signals in the LL_1 band into four additional subbands LL_2, LH_2, HL_2 and HH_2 and so on until a peak appears in the cross correlations. Otherwise, the signature can not be detected. The decoding method is illustrated in Fig. 5(b).

4. Numerical Examples

We implement two watermarking methods: one is using the DCT approach proposed by Cox el. al. in [2] and the other is using the DWT approach proposed in this paper. In the DWT approach, the Haar DWT is used. Two step DWT is implemented and images are decomposed into 7 subbands. Watermarks, Gaussian noise, are added into all 6 subbands but not in the lowest subband (the lowest frequency components). In the DCT approach, watermarks (Gaussian noise) are added to all the DCT coefficients. The levels of watermarks in the DWT and DCT approaches are the same, i.e., *the total energies of the watermark values in these two approaches are the same.* It should be noted that we have also implemented the DCT watermarking method when the pseudo-random sequence is added to the DCT values at the same positions as the ones in the above DWT approach, i.e., the middle frequencies. We found that the performance is not as good as the one by adding watermarks in all the frequencies in the DCT domain.

Two images with size 512×512, "peppers" and "car," are tested. Fig. 6(a) shows the original "peppers" image. Fig. 6(b) shows the watermarked image with the DWT approach and Fig. 7(a) shows the watermarked image with the DCT approach. Both watermarked images are indistinguishable from the original. A similar property holds for the second test image "car," whose original image is shown in Fig. 8(b).

The first distortion against which we test our algorithm with is additive noise. Two noisy images are shown in Fig. 7(b) and Fig. 8(a), respectively. When the variance of the additive noise is not too large, such as the one shown in Fig. 7(b), the signature can be detected only using the information in the HH_1 band with the DWT approach, where the cross correlations are shown in Fig. 9(a) and a peak can be clearly seen. When the variance of the additive noise is large, such as the one shown in Fig. 8(a), the HH_1 band information is not good enough with the DWT approach, where the cross correlations are shown in Fig. 9(b) and no clear peak can be seen. However, the signature can be detected by using the information in the HH_1 and LH_1 bands with the DWT approach, where the cross correlations are shown in Fig. 9(d) and a peak can be clearly seen. For the second noisy image, we have also implemented the DCT approach. In this case, the signature with the DCT approach can not be detected, where the correlations are shown in Fig. 9(c) and no clear peak can be seen. Similar results hold for the "car" image and the correlations are shown in Fig. 10.

The second "test" distortion is rescaling/stretching for "peppers" and "car" images. three types of rescaling/stretchings are implemented. In the first two implementations, the rescaled/stretched images are rescaled back to the same size of the original image using interpolations, where 25% reduction/enlargement is used. In the third implementation, the stretched images are simply cut back to the original size, where 1% and 2% stretching is used.

In the *rescaling*, an image, x, is reduced to 3/4 of the original size. The method of

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

the rescaling is from the MATLAB function called "imresize." as `imresize(x, 1-1/4, 'method')` where 'method' indicates one of the methods in the interpolations between pixels: piecewise constant, bilinear spline, and cubic spline. With the received smaller size image, for the watermark detection we extend it to the normal size, i.e., 512×512, by using the same Matlab function "imresize" as `imresize(y, 1+1/3, 'method')`, where 'method' is also one of the above interpolation methods. In this experiment, we implemented two different interpolation methods in `imresize` in the rescaling distortion: the piecewise constant method and the cubic spline method. In the detection, we alway use the cubic spline as `imresize(y, 1+1/3, 'bicubic')`. Similar results also hold for other combinations of these interpolation methods. Fig. 11 illustrate the detection results for the "peppers" image: Fig. 11(a),(c) show the cross correlations with the DWT approach while Fig. 11(b),(d) show the cross correlations with the DCT approach. In Fig. 11(a), (b), the rescaling method is `imresize(x,1-1/4,'nearest')`, i.e., the piecewise constant interpolation is used. In Fig. 11(c),(d), the rescaling method is `imresize(x,1-1/4,'bicubic')`, i.e., the cubic spline interpolation is used. One can see the better performance of the DWT approach over the DCT approach. Similar results hold for the "car" image and are shown in Fig. 12.

When, in the above rescaling experiment, the size of an image is first reduced and then extended in the detection, in the stretching, an image is first extended and then reduced in the detection. The same Matlab function `imresize` as in the rescaling is used. In the stretching experiment, an image is extended by 1/4 of the original size, i.e., the MATLAB function `imresize(x, 1+1/4, 'method')`, is used, where 'method' is the same as in the rescaling. In the detection, the received image is reduced by 1/5 to the original size, i.e., the Matlab function `imresize(y, 1-1/5, 'method')` is used. The rest is similar to the one in the rescaling. Figs. 13 and 14 show the correlation properties for the "peppers" and the "car" images, respectively.

In the third implementation of rescaling/stretching, an image is first stretched by 1% and 2% using the MATLAB function `imresize(y, 1+1/100, 'method')` and `imresize(y, 1+2/100, 'method')`, respectively. The stretched image is then cut back to the original size. Two images "peppers" and "car" are tested. Figs. 15-16 shows the correlation properties for the "peppers" and the "car" images, respectively, where (a) and (b) are for the 1% stretching, and (c) and (d) are for the 2% stretching.

The third "test" distortion is image compression. Two watermarked images with the DWT and DCT approaches shown in Fig. 6(b) and Fig. 7(a) are compressed by using the EZW coding algorithm. The compression ratio is chosen as 64, i.e., $0.125bpp$. With these two compressed images, the correlations are shown in Fig. 17 (a) and (b), where a peak in the middle can be clearly seen in Fig. 17(a) with the DWT approach, but no clear peaks can be seen in Fig. 17(b) with the DCT approach. This is not very surprising because the compression scheme is not suitable for the DCT approach. It should be noticed that the wavelet filters in the EZW compression are the commonly used Daubechies "9/7" biorthogonal wavelet filters while the wavelet filters in the watermarking are the simpliest Haar wavelet filters mentioned in Section 2.

The last "test" distortion is halftoning. The two watermarked images in Fig. 6(b) and Fig. 7(a) are both halftoned by using the following standard method. Let $x[m,n]$ be an image with 8 bit levels. To halftone it, we do the nonuniform thresholding through the Bayer's dither matrix T [17]:

$$T = (T_{j,k})_{4\times 4} = 16 \begin{pmatrix} 11 & 7 & 10 & 6 \\ 3 & 15 & 2 & 14 \\ 9 & 5 & 12 & 8 \\ 1 & 13 & 4 & 16 \end{pmatrix}$$

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

in the following way. Compare each disjoint 4×4 blocks in the image $x[m,n]$. If $x[m*4+j, n*4+k] \geq T_{j,k}$, then it is quantized to 1, and otherwise it is quantized to 0. Both DWT and DCT watermarking methods are tested. Surprisingly, we found that the watermarking method based on DWT we proposed in this paper is more robust than the method based on the DCT in [2-3]. The correlations are shown in Fig. 18(a) and (b), where (a) corresponds to the DWT approach while (b) corresponds to the DCT approach. One can clearly see a peak in the middle in Fig. 18(a) while no any clear peak in the middle can be seen in Fig. 18(b). In this experiment, the watermark was added to the middle frequencies in the DCT approach and no inverse halftoning was used.

5. Conclusion

In this paper, we have introduced a new multiresolution watermarking method using the discrete wavelet transform (DWT). In this method, Gaussian random noise is added to the large coefficients but not in the lowest subband in the DWT domain. The decoding is hierarchical. If distortion of a watermarked image is not serious, only a few bands worth of information are needed to detect the signature and therefore computational load can be saved. We have also implemented numerical examples for several kinds of distortions, such as additive noise, rescaling/stretching, compressed image with the wavelet approach such as the EZW, and halftoning. It is found that the DWT based watermark approach we proposed in this paper is robust to all the above distortions while the DCT approach is not, in particular, to distortions, such as compression, rescaling/stretching (1%, 2%, and 25% were tested), and additive noise with large noise variance.

6. Acknowledgements

Xia was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253 and the National Science Foundation CAREER Program under Grant MIP-9703377. Boncelet and Arce were supported in part through collaborative participation in the Advanced Telecommunications/Information Distribution Research Program (ATIRP) Consortium sponsored by the U.S. Army Research Laboratory under the Federated Laboratory Program, Cooperative Agreement DAAL01-96-0002. Arce was also supported in part by the National Science Foundation under the Grant MIP-9530923. They wish to thank the anonymous referees and the guest editor, Dr. Ingemar Cox, for their many helpful comments and suggestions that improved the clarity of this manuscript. They would also like to thank Mr. Jose Paredes for implementing numerous image compressions using the EZW method.

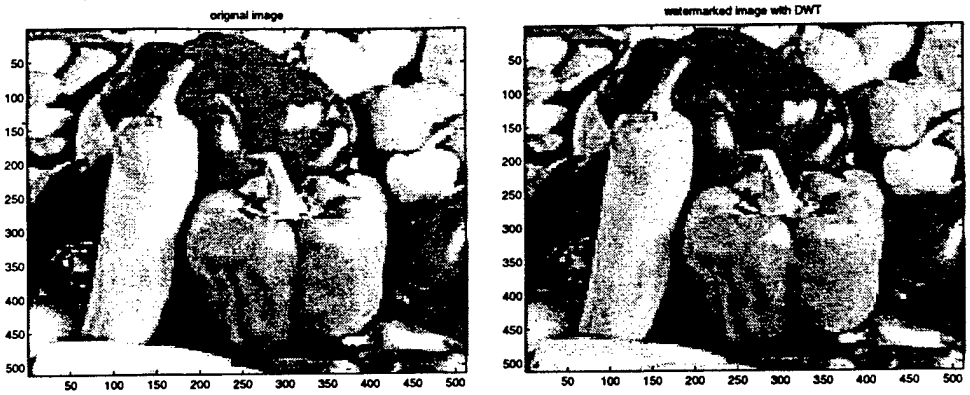

Figure 6. (a) Original "pepper" image; (b) Watermarked image using DWT.

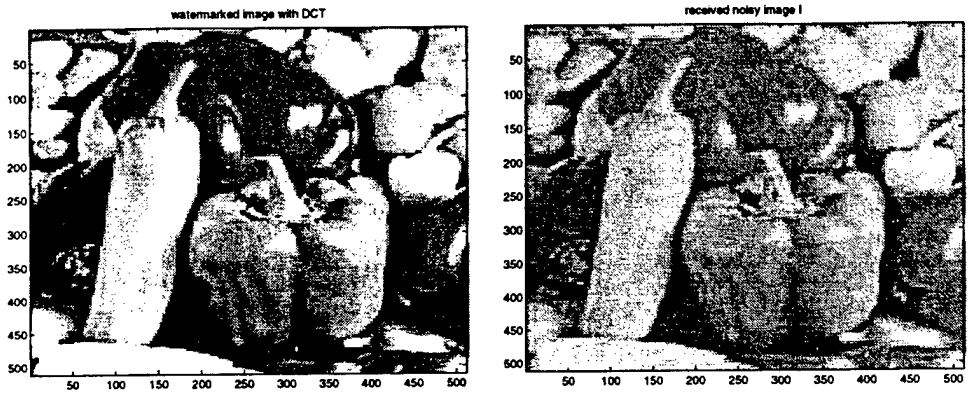

Figure 7. (a) Watermarked image using DCT; (b) Watermarked image with low additive noise.

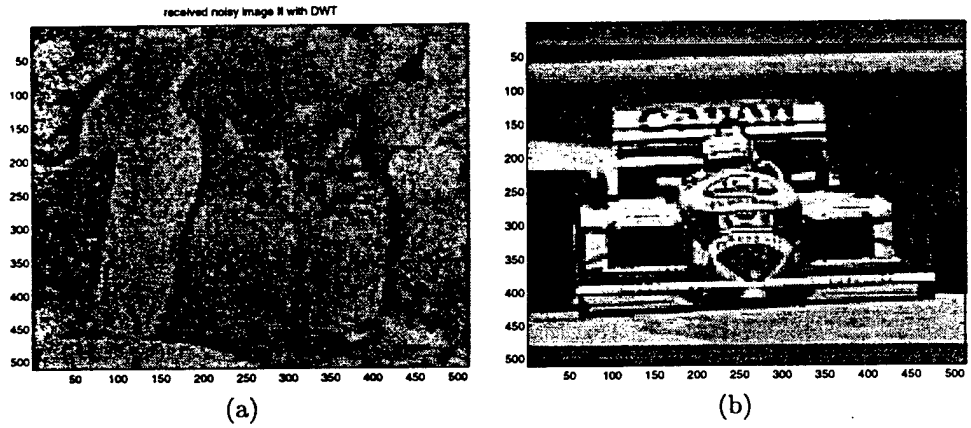

Figure 8. (a) Watermarked image with high additive noise; (b) Original "car" image.

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

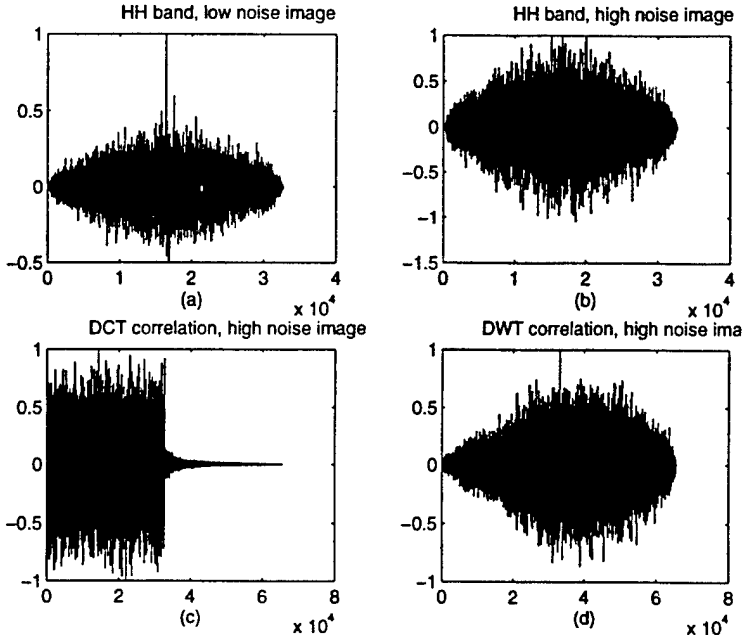

Figure 9. Correlations for watermark detection for the "peppers" image: (a) DWT with HH_1 band for low additive noise; (b) DWT with HH_1 band for high additive noise; (d) DWT with HH_1 and LH_1 bands for high additive noise; (c) DCT for high additive noise.

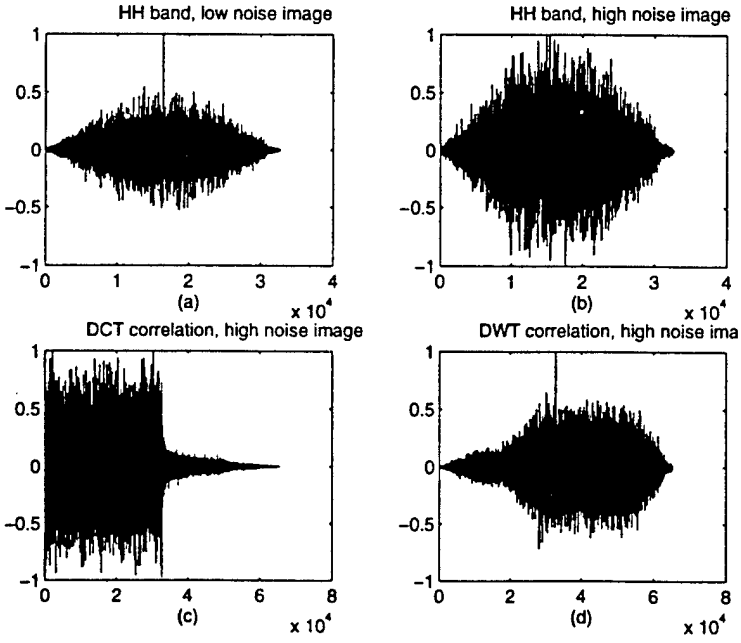

Figure 10. Correlations for watermark detection for the "car" image: (a) DWT with HH_1 band for low additive noise; (b) DWT with HH_1 band for high additive noise; (d) DWT with HH_1 and LH_1 bands for high additive noise; (c) DCT for high additive noise.

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

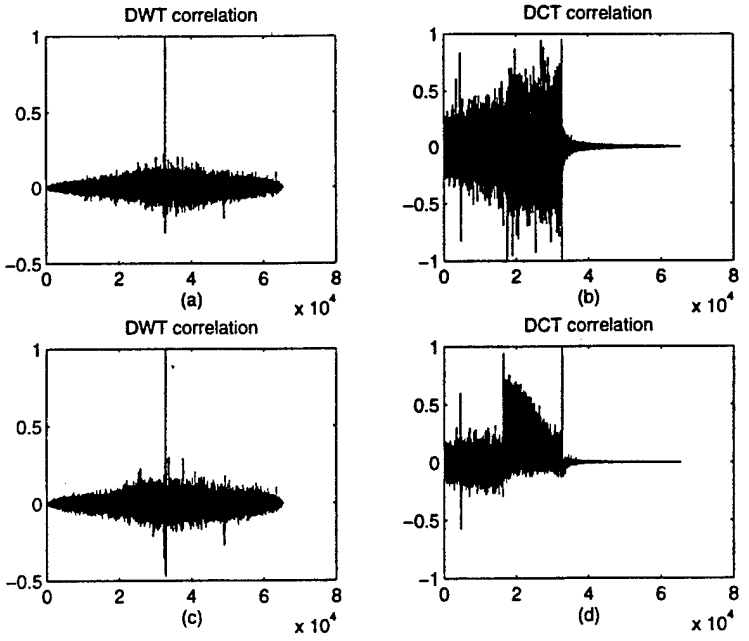

Figure 11. Correlations for watermark detection for the rescaled "peppers" image: (a) and (b) piecewise constant interpolation in the rescaling and (a) DWT (b) DCT; (c) and (d) cubic spline interpolation in the rescaling and (c) DWT (d) DCT.

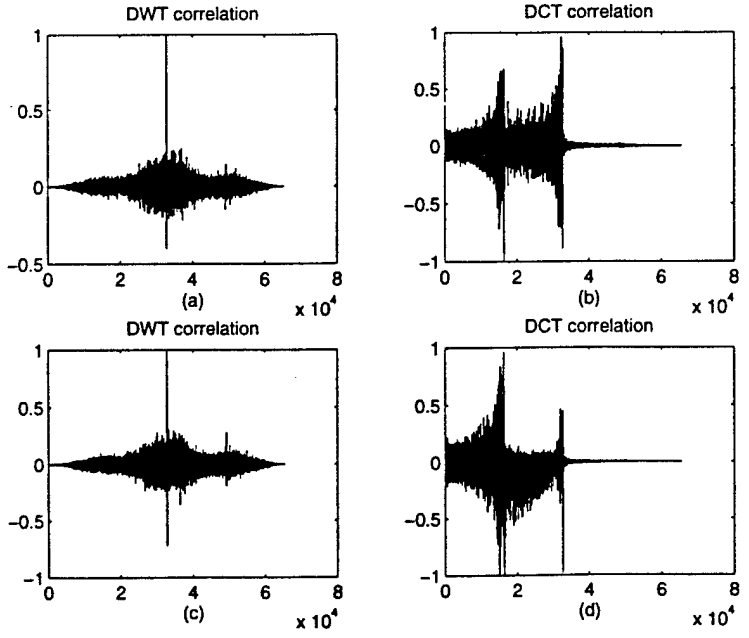

Figure 12. Correlations for watermark detection for the rescaled "car" image: (a) and (b) piecewise constant interpolation in the rescaling and (a) DWT (b) DCT; (c) and (d) cubic spline interpolation in the rescaling and (c) DWT (d) DCT.

#7038 - $15.00 US

Received October 14, 1998; Revised November 25, 1998

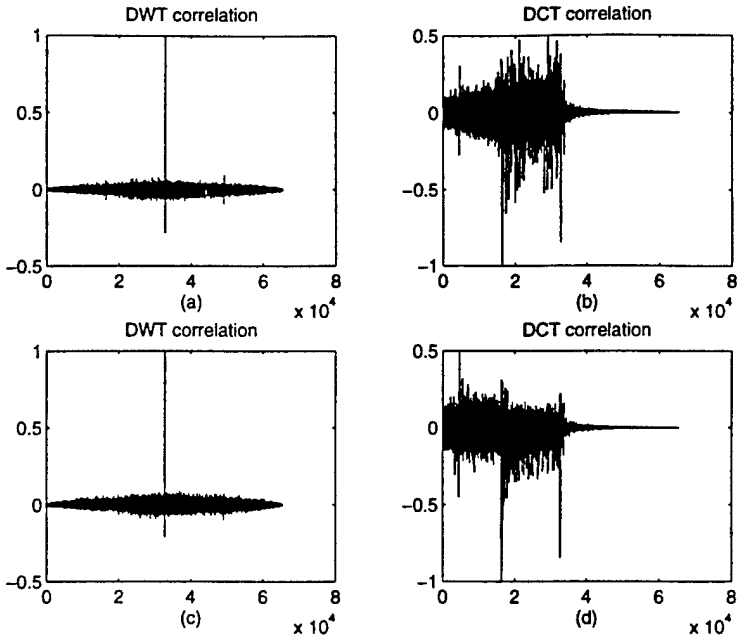

Figure 13. Correlations for watermark detection for the stretched "peppers" image: (a) and (b) piecewise constant interpolation in the rescaling and (a) DWT (b) DCT; (c) and (d) cubic spline interpolation in the rescaling and (c) DWT (d) DCT.

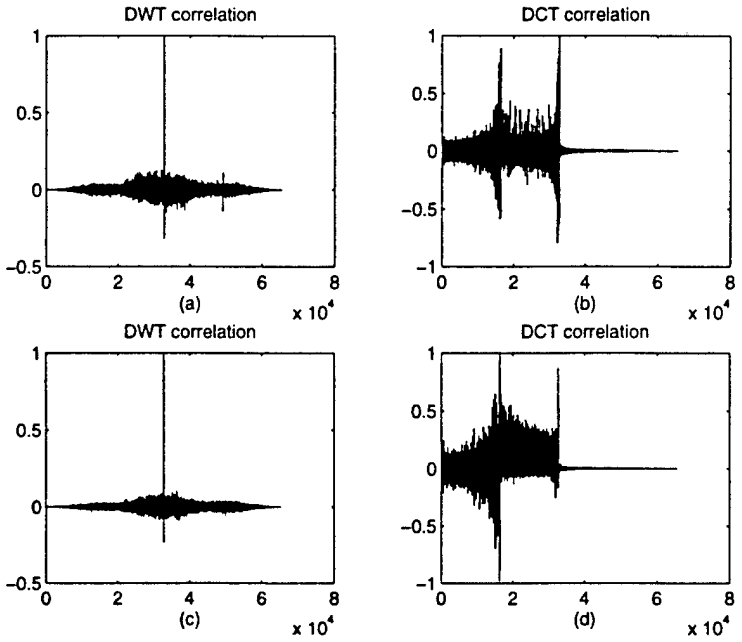

Figure 14. Correlations for watermark detection for the stretched "car" image: (a) and (b) piecewise constant interpolation in the rescaling and (a) DWT (b) DCT; (c) and (d) cubic spline interpolation in the rescaling and (c) DWT (d) DCT.

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

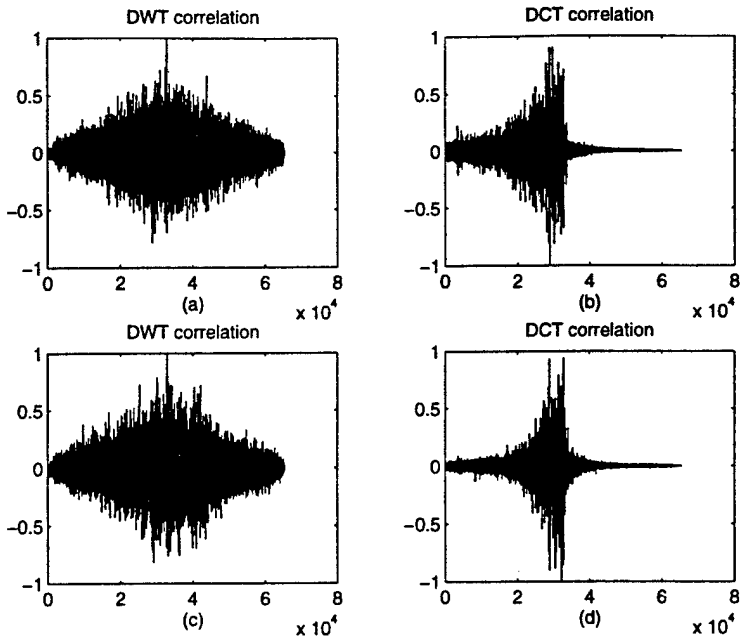

Figure 15. Correlations for watermark detection for the stretched "peppers" image: (a) and (b) 1% stretching and (a) DWT (b) DCT; (c) and (d) 2% stretching and (c) DWT (d) DCT.

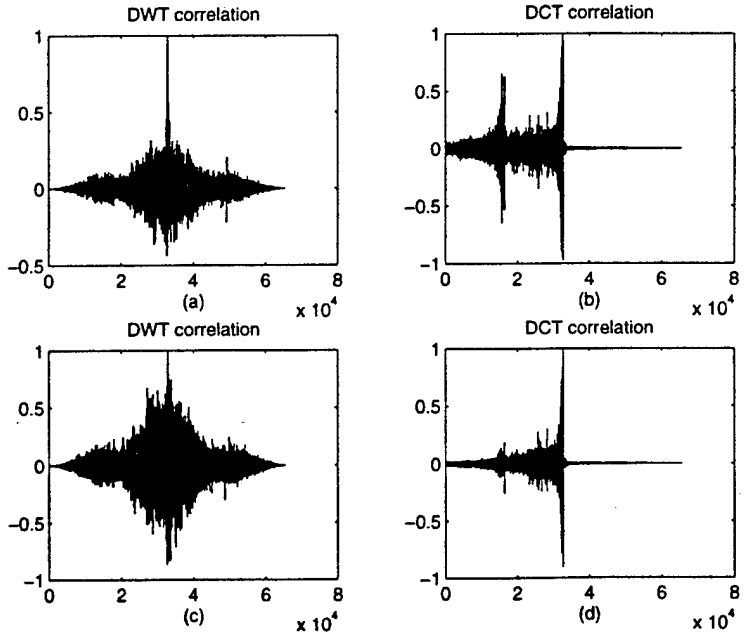

Figure 16. Correlations for watermark detection for the stretched "car" image: (a) and (b) 1% stretching and (a) DWT (b) DCT; (c) and (d) 2% stretching and (c) DWT (d) DCT.

#7038 - $15.00 US
Received October 14, 1998; Revised November 25, 1998

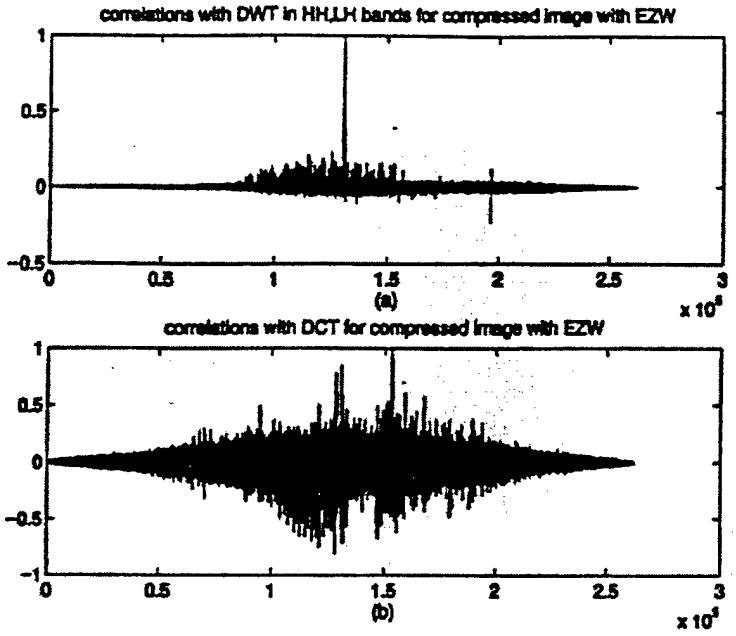

Figure 17. Correlations for watermark detection for compressed images: (a) DWT; (b) DCT.

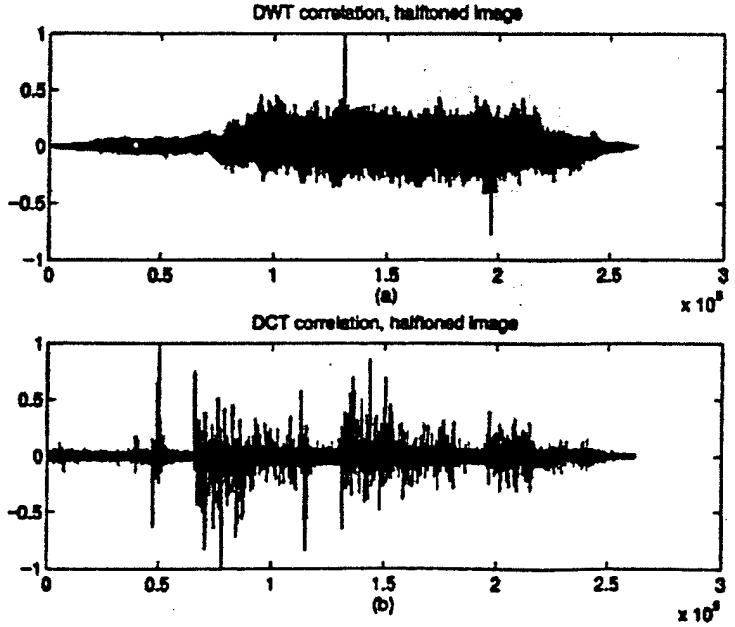

Figure 18. Correlations for watermark detection for halftoned images: (a) DWT; (b) DCT.

#7038 - $15.00 US

Received October 14, 1998; Revised November 25, 1998

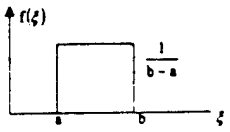

Fig 1 Uniform pdf.

PROOF The authors [1] assumed that both the target relative position and relative speed are uniformly distributed between 0 to 60 km and −40 to 40 m/s, respectively. Let us derive the variance of uniform distributed variable. The uniform probability density function (pdf) is defined by [2]

$$f(\xi) = \frac{1}{b-a} \qquad a \le \xi \le b$$
$$= 0 \qquad \text{otherwise}$$

for real constants $-\infty < a < \infty$ and $b > a$. Fig. 1 illustrates the behavior of the above function.

$$\sigma_\xi^2 = E[\xi^2] - \bar{\xi}^2$$
$$= \int_a^b \xi^2 \left(\frac{1}{b-a}\right) d\xi - \bar{\xi}^2$$
$$\sigma_\xi^2 = \frac{1}{b-a}\frac{\xi^3}{3}\Big|_a^b - \bar{\xi}^2$$
$$= \frac{1}{(b-a)}\frac{b^3-a^3}{3} - \left(\frac{b+a}{2}\right)^2$$
$$= \frac{(b-a)^2}{12}.$$

In case of relative range, x component $b = 60\sin\beta_0$, $a = 0$ and y component $b = 60\cos\beta_0$, $a = 0$. Hence

$$\sigma_{R_x}^2 = \frac{(60\sin\beta_0)^2}{12}, \qquad \sigma_{R_y}^2 = \frac{(60\cos\beta_0)^2}{12}.$$

In case of relative speed, $a = -40$ $b = 40$

$$\sigma_{S_x}^2 = \sigma_{S_y}^2 = \frac{(40+40)^2}{12} = \frac{80^2}{12}$$

$$P(0\,;0) = \text{diag}\left[\frac{(60\sin\beta_0)^2}{12} \quad \frac{80^2}{12} \quad \frac{(60\cos\beta_0)^2}{12} \quad \frac{80^2}{12}\right]$$
$$= \text{diag}\left[\frac{(60\sin\beta_0)^2}{12} \quad \frac{40^2}{3} \quad \frac{(60\cos\beta_0)^2}{12} \quad \frac{40^2}{3}\right].$$

S. KOTESWARA RAO
Scientist "F"
N.S.T.L.
Visakhapatnam-27, A.P.
India

REFERENCES

[1] Chan, Y. T., and Rudnicki, S. W. (1992)
Bearings-Only and Doppler-bearing tracking using instrumental variables.
IEEE Transactions on Aerospace and Electronic Systems, **28**, 4 (Oct. 1992), 1076–1082.

[2] Peebles, P. Z., Jr. (1987)
Probability, Random Variables and Random Signal Processing
New York: McGraw-Hill International, 1987.

Doppler Ambiguity Resolution Using Optimal Multiple Pulse Repetition Frequencies

Ferrari, Bérenguer, and Alengrin recently proposed an algorithm for velocity ambiguity resolution in coherent pulsed Doppler radar using multiple pulse repetition frequencies (PRFs). In this algorithm, two step estimations (folded frequency and ambiguity order) for the Doppler frequency is used by choosing particular PRF values. The folded frequency is the fractional part of the Doppler frequency and is estimated by averaging the folded frequency estimates for each PRF. The ambiguity order is the integer part of the Doppler frequency and is estimated by using the quasi-maximum-likelihood criterion. The PRF are grouped into pairs and each pair PRF values are symmetric about 1. The folded frequency estimate for each pair is the circular mean of the two folded frequency estimates of the pair due to the symmetry property.

We propose a new algorithm based on the optimal choice of the PRF values, where the PRF values are also grouped into pairs. In each pair PRF values, one is given and the other is optimally chosen. The optimality is built upon the minimal sidelobes of the maximum likelihood criterion. Numerical simulations are presented to illustrate the improved performance.

I. INTRODUCTION

Multiple pulse repetition frequency (PRF) is commonly used in modern-day radars for the velocity ambiguity resolution in coherent pulsed Doppler radars, see for example [1–4]. In this approach, the conventional method for achieving the ambiguity resolution is to search for the coincidence between unfolded Doppler frequency estimates for each PRF, see for example [2–4]. Since the Doppler frequency may take all possible real values in a range and infinite many trials are needed for all

Manuscript received September 10, 1977; revised March 1 and August 7, 1998.

IEEE Log No. T-AES/35/1/01518.

This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant F49620-97-1-0253, the National Science Foundation CAREER Program under Grant MIP-9703377, and the Office of Naval Research YIP under Grant N00014-98-1-0644.

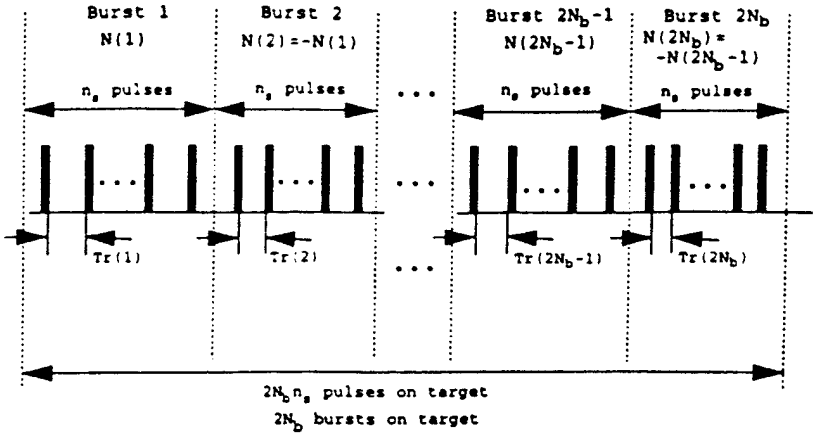

Fig. 1. Multiple PRF waveform.

the possibilities of the Doppler frequency, it maybe impossible to have an exact match. Thus, estimation errors usually occur. Based on this observation, a two step estimation algorithm has been proposed in [1] by Ferrari, Bérenguer, and Alengrin. The basic idea for the two step estimation is the following. The Doppler frequency is decomposed into two parts: the folded part, i.e., the fractional part modulo 1, and the ambiguity order part, i.e., the integer part. By grouping the PRFs into pairs where each pair is symmetric about 1, the folded part is the "circular mean" [5] of the folded estimates of the pair PRFs. This circular averaging is the first step of the algorithm in [1]. After the folded part is estimated, the second step is to find the match of the ambiguity order. By noticing that the ambiguity order takes integer values, there are only finite many possible trials needed ranging from the minimal and the maximal possible ambiguity orders. Therefore, the exact estimation of the Doppler frequency becomes possible with the two step estimation. Note that the *key* of this method is to convert the infinite many trials to the finite many trials, by converting a general real number matching to an integer matching.

The *motivation* for this paper is as follows. Since the specific PRF pairs, which are symmetric about 1, are needed in the Ferrari–Bérenguer–Alengrin approach, it may reduce the detectability of using the maximum likelihood criterion to detect the peak or the match. It is because the sidelobes of the maximum likelihood function with the specific PRFs may not be as low as the one with other PRFs. The motivation of this work is to relax the above PRF condition in the following way: one of each pair PRFs is fixed and the other of the pair is optimally determined based on the lowest sidelobes of the maximum likelihood function. With this relaxization, the "circular mean" estimation of the folded frequency may not be as good as the one in [1]. We propose an alternative approach to achieve the folded frequency estimation as follows. We first take the conventional mean of the folded frequency estimates in each pair. The true folded frequency falls in a finite number of possibilities from the conventional mean. These finite possibilities of the folded frequency can be obtained when the PRF pairs are known. Since the ambiguity order has also finite possibilities, the overall folded frequency and the ambiguity order have *finite* possibilities. This suggests us to estimate both the folded frequency and the ambiguity order simultaneously based on the maximum likelihood criterion. What is gained here is the detectability improvement of the Doppler frequency while the penalty is the increase of the computational complexity with a multiple of the one in [1] due to more possibilities to search for the folded frequency. .

This paper is organized as follows. In Section II, we briefly review the Ferrari–Bérenguer–Alengrin approach proposed in [1]. In Section III, we study the optimal PRF method. In Section IV, we present numerical examples which outperform the Ferrari–Bérenguer–Alengrin method.

II. THE FERRARI–BÉRENGUER–ALENGRIN TWO STEP ESTIMATION METHOD

First of all, we briefly describe the problem. Let radar transmit $2N_b$ bursts of n_s pulses, where the PRF of the kth burst is assumed $Fr(k)$, $1 \le k \le 2N_b$. The time difference between two pulses in the kth burst is $Tr(k) = 1/Fr(k)$. It is assumed that the elapsed time between the last pulse of the kth burst and the first one of the $(k+1)$th burst is $Tr(k)$. The time delays $Tr(k)$ are assumed as

$$Tr(k) = \left(1 + \frac{1}{N(k)}\right) Tr \tag{1}$$

where $N(1), N(2), \ldots, N(2N_b)$ are integers and Tr is usually assumed as 1 for simplicity. The multiple PRF waveform is shown in Fig. 1.

After coherent demodulation, the received data at the nth sample, $0 \le n \le n_s - 1$, in the kth burst,

$1 \le k \le 2N_b$, becomes

$$y_k(n) = x_k(n) + b_k(n) = a_k(f_D)\exp(j2\pi f_D nTr(k)) + b_k(n) \tag{2}$$

where f_D is the unknown Doppler frequency, $b_k(n)$ is white noise from the contribution of both thermal noise and clutter whitened residue, and $a_k(f_D)$ contains the initial phase of the target signal on the kth burst. If $a_1(f_D) = A$, then we have

$$a_k(f_D) = A\exp\left(j2\pi n_s f_D \sum_{q=1}^{k-1} Tr(q)\right), \qquad k \ge 2. \tag{3}$$

Then the ambiguity resolution problem is to estimate the Doppler frequency f_D from the noisy data $y_k(n)$ in (2). It is usually assumed that f_D is in a certain range, i.e., $|f| \le f_{\max}$. The conventional detection method is the following maximum likelihood estimation. Find $\hat{f}_D$ that maximizes the following maximum likelihood function

$$L(f) = \left|\sum_{k=1}^{2N_b}\sum_{m=0}^{n_s-1} y_k(m)a_k^*(f)\exp(-j2\pi f mTr(k))\right|^2 \tag{4}$$

i.e.,

$$L(\hat{f}_D) = \max_{f:\le f_{\max}} L(f)$$

where $a_k(f)$ takes the form (3) with f_D replaced by f and $|\hat{f}_D| \le f_{\max}$. This is a matching process and f needs to run all real numbers from $-f_{\max}$ to $f_{\max}$. Clearly, it has infinite many trials and therefore is impossible to have an exact match.

In [1], Ferrari–Bérenguer–Alengrin proposed an alternative two step approach for the above problem without implementing infinite many trials, where particular $N(k)$ in (1) were used. We next want to briefly describe this two step approach.

Let $N(2p+1)$ be a positive integer and set

$$N(2p+2) = -N(2p+1), \qquad \text{for} \quad p = 0,1,\ldots,N_b - 1. \tag{5}$$

The Doppler frequency f_D is decomposed into its integer part (the ambiguity order) n_r and fractional part (the folded or reduced frequency) f_r as

$$f_D = f_r + n_r \qquad \text{with} \quad 0 \le f_r < 1. \tag{6}$$

Then (2) becomes

$$y_k(n) = a_k(f_D)\exp\left(j2\pi\left(f_r + \frac{f_D}{N(k)}\right)n\right) + b_k(n), \qquad 0 \le n \le n_s - 1. \tag{7}$$

Let

$$f_k \triangleq f_r + \frac{f_D}{N(k)}, \qquad 1 \le k \le 2N_b. \tag{8}$$

If f_k could be obtained from $y_k(n)$, $0 \le n \le n_s - 1$, in (7), by using $N(2p+2) = -N(2p+1)$ in (5), the reduced frequency f_r would be

$$f_r = \frac{f_{2p+1} + f_{2p+2}}{2}, \qquad 0 \le p \le N_b - 1. \tag{9}$$

From $y_k(n)$ in (7) what we can get for f_k is, however, its folded version $\tilde{f}_k$, i.e.,

$$\tilde{f}_k = f_k + l, \qquad l \text{ is an unknown integer and } 0 \le \tilde{f}_k < 1. \tag{10}$$

In this case, the reduced frequency f_r cannot be obtained from $\tilde{f}_k$ by simply taking their mean as $(\tilde{f}_{2p+1} + \tilde{f}_{2p+2})/2$. However, when

$$|f_{2p+1} - f_{2p+2}| < 0.5 \tag{11}$$

the reduced frequency f_r can be recovered from $\tilde{f}_k$ by taking the "circular mean" [5] as

$$\hat{f}_r(p) = \frac{1}{2\pi}\text{angle}[\exp(j2\pi\tilde{f}_{2p+1}) + \exp(j2\pi\tilde{f}_{2p+2})] \tag{12}$$

where angle(z) is the phase angle in radians in $[0, 2\pi)$ of the complex number z. With total N_b pairs of f_k, the overall estimate of the reduced frequency f_r is

$$\hat{f}_r = \frac{1}{2\pi}\text{angle}\sum_{p=0}^{N_b-1}\exp(j2\pi\hat{f}_r(p)). \tag{13}$$

When the Doppler frequency f in (4) is split into its reduced frequency part f (without confusion in understanding we also use f to denote the reduced frequency) and its ambiguity order part n, the maximum likelihood function in (4) can be written as

$$L(f,n) \triangleq \left|\sum_{k=1}^{2N_b}\sum_{m=0}^{n_s-1} y_k(m)a_k^*(f,n) \cdot \exp(-j2\pi(f+n)mTr(k))\right|^2$$

$$= \left|\sum_{k=1}^{2N_b} a_k^*(f,n)\sum_{m=0}^{n_s-1} y_k(m)\exp(-j2\pi f mTr(k)) \cdot \exp\left(-j2\pi\frac{n}{N(k)}m\right)\right|^2 \tag{14}$$

where $a_k(f,n)$ corresponds to the term $a_k(f)$ in (4) and can be expressed as

$$a_{2p+1}(f,n) = \exp(j2\pi f n_s 2p) \tag{15}$$

$$a_{2p+2}(f,n) = \exp\left(j2\pi f n_s\left(2p+1+\frac{1}{N(2p+1)}\right)\right) \cdot \exp\left(j2\pi n_s \frac{n}{N(2p+1)}\right). \tag{16}$$

After the reduced frequency f_r is estimated as in (13), the maximum likelihood function $L(f,n)$ for both f and n is reduced to the one for the ambiguity order n only:

$$L(n) \triangleq L(\hat{f}_r,n) = \sum_{k=1}^{2N_b} a_k^*(\hat{f}_r,n)\sum_{m=0}^{n_s-1} y_k(m)\exp(-j2\pi \hat{f}_r m T r(k)) \cdot \exp\left(-j2\pi\frac{n}{N(k)}m\right)^2 \tag{17}$$

where n ranges all integers from $-n_{\max}$ to $n_{\max}$ and $n_{\max}$ is the maximum ambiguity order corresponding to the maximum Doppler frequency $f_{\max}$. Thus, the searching of the Doppler frequency f_D from all the real numbers $|f| \le f_{\max}$ to maximize $L(f)$ in (4) becomes the searching of the ambiguity order n from all integers $|n| \le n_{\max}$ to maximize $L(n)$ in (17). Note that there are only finite many possibilities of n, which makes the exact coincidence of the true ambiguity order possible. Let $\hat{n}_r$ denote the optimal ambiguity order estimate from $L(n)$ in (17). Then the final Doppler frequency estimate is

$$\hat{f}_D = \hat{f}_r + \hat{n}_r. \tag{18}$$

The reason for choosing $N(k)$ as integers in the whole approach is to use the discrete Fourier transform (DFT) calculations in (17) for the maximum likelihood function evaluations. For more details on the implementation issue, see [1].

The above is the main idea for the Ferrari–Bérenguer–Alengrin two step estimation method. We call it *FBA method*. It is built upon the assumption (5) and the condition (11). Condition (11) guarantees the accurate reduced frequency estimation and leads to the following condition on $N(k)$:

$$|N(k)| > 4(1+n_{\max}), \qquad 1 \le k \le 2N_b \tag{19}$$

where $n_{\max}$ is the maximum ambiguity order. Clearly, when $n_{\max}$ is large, $|N(k)|$ needs to be large. Large $|N(k)|$ may increase ambiguity order errors as mentioned in [1]. One way to relax the condition (11) or (19) is as follows, which also serves as a foundation for the optimal multiple PRF discussed latter.

Assume

$$\frac{f_D}{N(k)} < 1, \quad \text{i.e.,} \quad |N(k)| > 1 + n_{\max}. \tag{20}$$

In this case, although the circular mean (12) may not be equal to the reduced frequency f_r in (8), f_r takes one of the following five values:

$$\overline{f}_r(p), \quad \overline{f}_r(p)-0.5, \quad \overline{f}_r(p)+0.5, \quad \overline{f}_r(p)-1, \quad \overline{f}_r(p)+1 \tag{21}$$

where $\overline{f}_r(p)$ is the conventional mean, $\overline{f}_r(p) = (\tilde{f}_{2p+1} + \tilde{f}_{2p+2})/2$, and $\tilde{f}_k$ are obtained from (7) and (10). It is because the unknown parameter l in (10) may only take 0, -1 or 1, when the condition (20) holds and $0 \le f_r < 1$. Thus, when $N_b = 1$, the estimation of f_r and n_r become the search of the optimal $\hat{f}_r(p)$ and $\hat{n}_r$ in the maximum likelehood function $L(f_r,n)$ in (14) among

$$f_r \in S(p) \triangleq \{\overline{f}_r(p), \overline{f}_r(p)-0.5, \overline{f}_r(p)+0.5, \overline{f}_r(p)-1, \overline{f}_r(p)+1\} \tag{22}$$

and $|n| \le n_{\max}$:

$$L(\hat{f}_r(p),\hat{n}_r) = \max_{f\in S(p),|n|\le n_{\max}} L(f,n) \tag{23}$$

which also has only finite comparisons.

When $N_b > 1$, there are at least two methods to take the advantage of this multiplicity. One is to take the circular mean of all the above estimated $\hat{f}_r(p)$ as in (13). The other is to search the optimal f among all possible elements in the sets $S(p)$ for $p = 0,1,\ldots,N_b-1$:

$$L(\hat{f}_r,\hat{n}_r) = \max_{f\in S,|n|\le n_{\max}} L(f,n) \tag{24}$$

where

$$S = \bigcup_{p=0}^{N_b-1} S(p).$$

Note that the condition (20) can be further relaxed by allowing more possibilities for the reduced frequency f_r from the mean $\overline{f}_r$. Thus, the size of $N(k)$ can basically be arbitrary. The detection method in (20)–(24) is called *modified FBA method*. On the other hand, the condition (5) may cause high sidelobes of the maximum likelihood function $L(f_r,n)$ in (14) and therefore reduce the performance when additive white noise $b_k(n)$ in (2) is significant. The goal of the rest of this paper is to relax the condition (5) and search for the optimal linear relationship between $N(2p+1)$ and $N(2p+2)$ instead of $N(2p+2) = -N(2p+1)$.

It should be mentioned that another difference between the FBA method and the above modified FBA method is the following. In the FBA method, the angular mean is taken over the N_b bursts as shown

in (13), while, in the modified FBA method, the multiplicity of the bursts gives more possibilities to search for the correct folded frequency. The angular mean may reduce the error variance of the reduced frequency, while the more possibilities of the search may provide more accurate estimate of the reduced frequency. However, the latter one clearly causes more computations.

III. OPTIMAL MULTIPLE PRF AND DOPPLER FREQUENCY DETECTION

In this section, we use the same signal model as described in Section II, where the assumption (5) is relaxed as

$$N(2p+2) = -\alpha_p N(2p+1), \quad \text{for} \quad p = 0,1,\ldots,N_b-1 \tag{25}$$

where $N(2p+1)$ are positive integers and α_p are positive real parameters. The goal of the rest of this paper is to optimally determine the parameters α_p given $N(2p+1)$ for $p = 0,1,\ldots,N_b-1$ in terms of the lowest sidelobes of the maximum likelihood function $L(f,n)$.

With (25), an analogous formula of (9) for the reduced frequency is

$$f_r = \frac{f_{2p+1} + \alpha_p f_{2p+2}}{1+\alpha_p}, \qquad p = 0,1,\ldots,N_b-1 \tag{26}$$

where f_k are defined in (8). One can see that the conventional mean (9) with the property (5) becomes the conventional weighted mean (26) with the property (25). The circular mean in (12), however, cannot be generalized to the general setting of the parameters α_p. In other words, the reduced frequency f_r can not be obtained as in the FBA method from the estimated individual $\tilde{f}_k$ in (8), (10), and (25) with general parameters α_p unless $\alpha_p = 1$ using the periodogram method. Fortunately, the argument in (20)–(24) can be generalized as follows.

Without loss of generality, we assume the property (20), i.e.,

$$N(2p+1) > 1 + n_{\max} \quad \text{and} \quad |N(2p+1)| > \frac{1+n_{\max}}{\alpha_p}, \qquad p = 0,1,\ldots,N_b-1. \tag{27}$$

Let

$$\overline{f}_r(p) \triangleq \frac{\tilde{f}_{2p+1} + \alpha_p \tilde{f}_{2p+2}}{1+\alpha_p}, \qquad p = 0,1,\ldots,N_b-1 \tag{28}$$

where $\tilde{f}_k$ are obtained from (7), (8), and (10) with $N(k)$ satisfying (25) instead of (5). For $p = 0,1,\ldots,N_b-1$, let

$$S(p) \triangleq \left\{ \overline{f}_r(p), \overline{f}_r(p) \pm \frac{1}{1+\alpha_p}, \overline{f}_r(p) \pm \frac{\alpha_p}{1+\alpha_p}, \overline{f}_r(p) \pm \frac{1-\alpha_p}{1+\alpha_p}, \overline{f}_r(p) \pm 1 \right\}. \tag{29}$$

When $\alpha_p = 1$, the set $S(p)$ in (29) is the same as the set $S(p)$ in (22). Similar to (21), we have

$$f_r \in S(p), \qquad p = 0,1,\ldots,N_b-1. \tag{30}$$

Let

$$S = \bigcup_{p=0}^{N_b-1} S(p). \tag{31}$$

Then the maximum likelihood estimates for the reduced frequency f_r and the ambiguity order n_r are $\hat{f}_r$ and $\hat{n}_r$ that maximize $L(f,n)$ for $f \in S$ and $|n| \le n_{\max}$, i.e.,

$$L(\hat{f}_r, \hat{n}_r) = \max_{f \in S, |n| \le n_{\max}} L(f,n) \tag{32}$$

where $L(f,n)$ is similar to (14):

$$L(f,n) = \left| \sum_{k=1}^{2N_b} a_k^*(f,n) \sum_{m=0}^{n_s-1} y_k(m) \exp(-j2\pi f m Tr(k)) \cdot \exp\left(-j2\pi \frac{n}{N(k)} m\right) \right|^2 \tag{33}$$

where

$$a_k(f,n) = A \exp\left(j2\pi n_s (f+n) \sum_{q=1}^{k-1} Tr(q) \right), \qquad Tr(q) = 1 + \frac{1}{N(q)} \tag{34}$$

and $y_k(m)$ are the demodulated noisy data at the receiver:

$$y_k(m) = a_k(f_r, n_r) \exp(j2\pi f_r m Tr(k)) \cdot \exp\left(j2\pi \frac{n_r}{N(k)} m\right) + b_k(m) \tag{35}$$

where $f_D = f_r + n_r$ is the unknown Doppler frequency and $b_k(m)$ are additive white noise. The final Doppler frequency estimate is $\hat{f}_D = \hat{f}_r + \hat{n}_r$.

The performance of the above detection method depends on the property of the maximum likelihood function $L(f,n)$. The lower sidelobes of $L(f,n)$ are, the better performance of the detection is. The sidelobes depend on the choice of the parameters α_p in (35), when $N(2p+1)$ are given. We next want to discuss the optimal choice of these parameters.

By substituting (34)–(35) into (33), we have

$$L(f,n) = |A|^2 \left| \sum_{k=1}^{2N_b} \exp\left(j2\pi n_s (f_r - f + n_r - n) \sum_{q=1}^{k-1} Tr(q) \right) \cdot \sum_{m=0}^{n_r - 1} \exp(j2\pi (f_r - f) m Tr(k)) \exp\left(j2\pi \frac{n_r - n}{N(k)} m \right) \right|^2$$

$$= |A|^2 \left| \sum_{k=1}^{2N_b} \exp\left(j2\pi n_s (f_r - f + n_r - n) \sum_{q=1}^{k-1} Tr(q) \right) \cdot \exp\left(j\pi (n_s - 1) \left((f_r - f) Tr(k) + \frac{n_r - n}{N(k)} \right) \right) \cdot \frac{\sin\left(\pi n_s \left[(f_r - f) Tr(k) + \frac{n_r - n}{N(k)} \right] \right)}{\sin\left(\pi \left[(f_r - f) Tr(k) + \frac{n_r - n}{N(k)} \right] \right)} \right|^2 \quad (36)$$

Clearly, the mainlobe value of the above maximum likelihood function is its value when $f = f_r$ and $n = n_r$:

$$L(f_r, n_r) = |A|^2 2N_b n_s. \quad (37)$$

Since $f_r \in \mathcal{S}(p)$, the offset value $f_r - f$ in (36) may only take the values in the following set, when $f \in S$ defined in (30):

$$S_{\text{offset}} \triangleq \bigcup_{p=0}^{N_b - 1} \left\{ \pm 1, \pm 2, \frac{\pm 1}{1+\alpha_p}, \frac{\pm 2}{1+\alpha_p}, \frac{\pm \alpha_p}{1+\alpha_p}, \frac{\pm 2\alpha_p}{1+\alpha_p}, \frac{\pm (1-\alpha_p)}{1+\alpha_p}, \frac{\pm 2(1-\alpha_p)}{1+\alpha_p}, \frac{\pm (1 \pm 2\alpha_p)}{1+\alpha_p}, \frac{\pm (2 \pm \alpha_p)}{1+\alpha_p} \right\}. \quad (38)$$

The offset value $n_r - n$ is in the set $\{\pm 1, \pm 2, \ldots, \pm 2n_{\max}\}$.

Let $E_{\text{sidelobe}}(\alpha_0, \alpha_1, \ldots, \alpha_{N_b - 1})$ denote the total energy of all the sidelobe values of the maximum likelihood function $L(f,n)$ in (36). Then, by normalizing $A = 1$ it can be expressed by

$$E_{\text{sidelobe}}(\alpha_0, \alpha_1, \ldots, \alpha_{N_b - 1}) = \sum_{f \in S_{\text{offset}}} \sum_{0 < |n| \le 2n_{\max}} \left| \sum_{k=1}^{2N_b} \exp\left(j2\pi n_s (f+n) \sum_{q=1}^{k-1} \left(1 + \frac{1}{N(q)} \right) \right) \cdot \exp\left(j\pi (n_s - 1) \left(f\left(1 + \frac{1}{N(k)} \right) + \frac{n}{N(k)} \right) \right) \cdot \frac{\sin\left(\pi n_s \left[f\left(1 + \frac{1}{N(k)} \right) + \frac{n}{N(k)} \right] \right)}{\sin\left(\pi \left[f\left(1 + \frac{1}{N(k)} \right) + \frac{n}{N(k)} \right] \right)} \right|^2 \quad (39)$$

where $N(2p+2) = -\alpha_p N(2p+1)$, $p = 0, 1, \ldots, N_b - 1$, and S_{offset} is defined in (38). Given $N(2p+1)$, $p = 0, 1, \ldots, N_b - 1$, and $n_{\max}$, the optimal parameters $\hat{\alpha}_p$, $p = 0, 1, \ldots, N_b - 1$, can be obtained by minimizing the cost function $E_{\text{sidelobe}}(\alpha_0, \alpha_1, \ldots, \alpha_{N_b - 1})$ in (39) for $\hat{\alpha}_p > 0$, i.e.,

$$E_{\text{sidelobe}}(\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_{N_b - 1}) = \min_{\alpha_0 > a_0, \alpha_1 > a_1, \ldots, \alpha_{N_b - 1} > a_{N_b - 1}} E_{\text{sidelobe}}(\alpha_0, \alpha_1, \ldots, \alpha_{N_b - 1}) \quad (40)$$

where, by (27),

$$a_p = \frac{1 + n_{\max}}{N(2p+1)}.$$

One may see that an explicit solution for the optimal α_p is not possible. However, any existing optimization methods work for the above problem.

Let us consider the simpliest case, $N_b = 1$. In this case,

$$S_{\text{offset}} \triangleq \left\{ \pm 1, \pm 2, \frac{\pm 1}{1+\alpha_0}, \frac{\pm 2}{1+\alpha_0}, \frac{\pm \alpha_0}{1+\alpha_0}, \frac{\pm 2\alpha_0}{1+\alpha_0}, \frac{\pm (1-\alpha_0)}{1+\alpha_0}, \frac{\pm 2(1-\alpha_0)}{1+\alpha_0}, \frac{\pm (1 \pm 2\alpha_0)}{1+\alpha_0}, \frac{\pm (2 \pm \alpha_0)}{1+\alpha_0} \right\} \quad (41)$$

and

$$E_{\text{sidelobe}}(\alpha_0) = \sum_{f \in S_{\text{offset}}} \sum_{0 < |n| \le 2n_{\max}} \left| \frac{\sin\left(\pi n_s \left[f\left(1 + \frac{1}{N(1)} \right) + \frac{n}{N(1)} \right] \right)}{\sin\left(\pi \left[f\left(1 + \frac{1}{N(1)} \right) + \frac{n}{N(1)} \right] \right)} + \exp\left(j2\pi n_s (f+n) \left(1 + \frac{1}{N(1)} \right) \right) \cdot \exp\left(j\pi (n_s - 1) \left[f\left(1 - \frac{1}{\alpha_0 N(1)} \right) - \frac{n}{\alpha_0 N(1)} \right] \right) \cdot \frac{\sin\left(\pi n_s \left[f\left(1 - \frac{1}{\alpha_0 N(1)} \right) - \frac{n}{\alpha_0 N(1)} \right] \right)}{\sin\left(\pi \left[f\left(1 - \frac{1}{\alpha_0 N(1)} \right) - \frac{n}{\alpha_0 N(1)} \right] \right)} \right|^2. \quad (42)$$

Let us see some numerical examples of $E_{\text{sidelobe}}(\alpha_0)$. Consider $N(1) = 40$ and $n_s = 12$. Figs. 2, 3, and 4 show the $E_{\text{sidelobe}}(\alpha_0)$ versus α_0 when $n_{\max} = 3$, 5, and 12, respectively. One can see that the optimal α_0 strongly depends on the maximal ambiguity order $n_{\max}$, where the optimal α_0 are $\hat{\alpha}_0 = 0.57$, 1.85, and 2.01 for $n_{\max} = 3$, 5, and 12, respectively.

IV. NUMERICAL EXPERIMENTS

In this section, we present numerical examples to compare the performances for the modified FBA method and the method with optimized PRFs

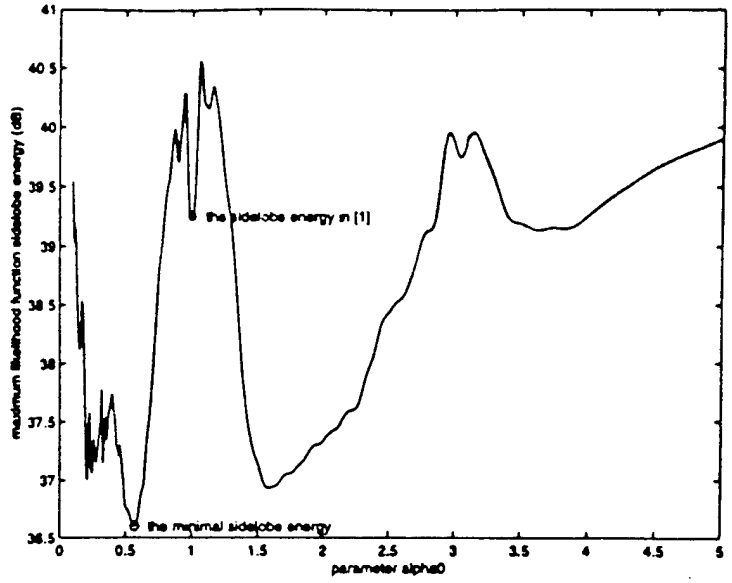

Fig. 2. $E_{\text{sidelobe}}(\alpha_0)$ when $N(1) = 40$, $n_s = 12$, and $n_{\max} = 3$. Optimal $\hat{\alpha}_0 = 0.57$.

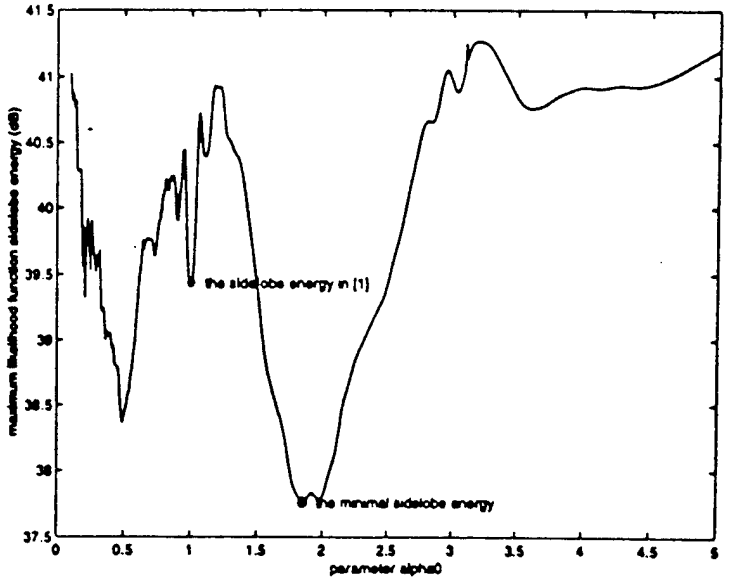

Fig. 3. $E_{\text{sidelobe}}(\alpha_0)$ when $N(1) = 40$, $n_s = 12$, and $n_{\max} = 5$. Optimal $\hat{\alpha}_0 = 1.85$.

proposed in this work. The following parameters are used in our simulations: $N(1) = 40$, $N_b = 1$, $n_s = 12$, and $N(2) = -\alpha_0 N(1)$, where $\alpha_0 = 1$ for FBA method and α_0 the optimal $\hat{\alpha}_0$ for the method proposed in this work. The additive noise $b_k(n)$ in the known noisy radar data $y_k(n)$ in (2) is assumed white Gaussian noise with mean 0 and variance σ^2. As mentioned at the end of Section III, the optimal α_0 depends on the maximal ambiguity order $n_{\max}$. Two different $n_{\max}$ are tested: $n_{\max} = 3$ and 12. Let M be the number of signal realizations. Let $f_D(k)$ be the true Doppler frequency and $\hat{f}_D(k)$ be the estimated one at the kth signal realization. Then the mean squared error (MSE) is calculated as

$$\text{MSE} = \frac{\sum_{k=1}^{M} |\hat{f}_D(k) - f_D(k)|^2}{M}. \tag{43}$$

The signal-to-noise ratio (SNR) for the additive Gaussian noise is calculated by $\text{SNR} = A^2/\sigma^2$, where A is the transmitted signal amplitude.

When $n_{\max} = 3$ and $N(1) = -N(2) = 40 > 4(1 + 3) = 16$, i.e., the condition (19) or (11) holds for the accurate circular mean formula (12). The FBA method works in this case although the parameter $\alpha = 1$ is not optimal in terms of the sidelobe values of the maximum likelihood function $E_{\text{sidelobe}}(\alpha_0)$. The optimal parameter α_0 in this case is $\hat{\alpha}_0 = 0.57$ as studied in Section III. When $\alpha_0 = 0.57$, clearly the number $N(2) = -\alpha_0 N(1) = 22.8$ is not an integer. For the DFT computation purpose, rounding $\alpha = 0.57$ to $\alpha_0 = 0.6$ may be needed for $N(2)$ to be an integer. When $\alpha_0 = 0.6$, $N(2) = -24$. As mentioned in Section III, when $\alpha_0 \neq 1$, the accurate circular mean no longer

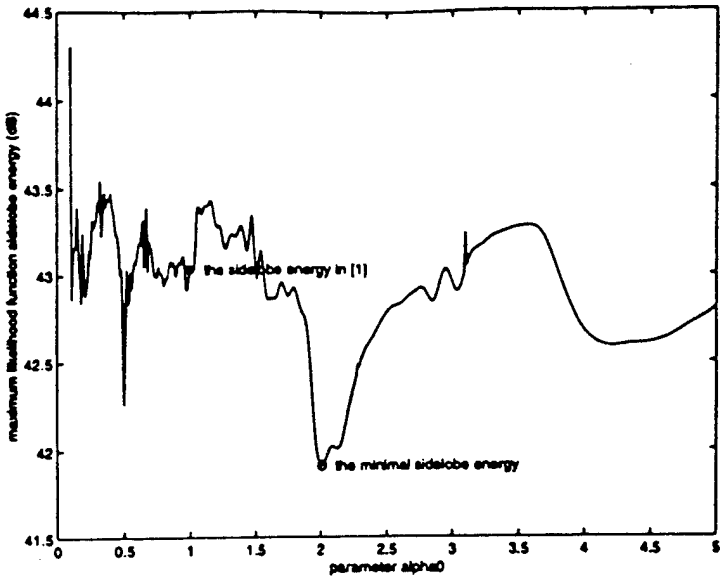

Fig. 4. $E_{\text{sidelobe}}(\alpha_0)$ when $N(1) = 40$, $n_s = 12$, and $n_{\max} = 12$. Optimal $\hat{\alpha}_0 = 2.01$.

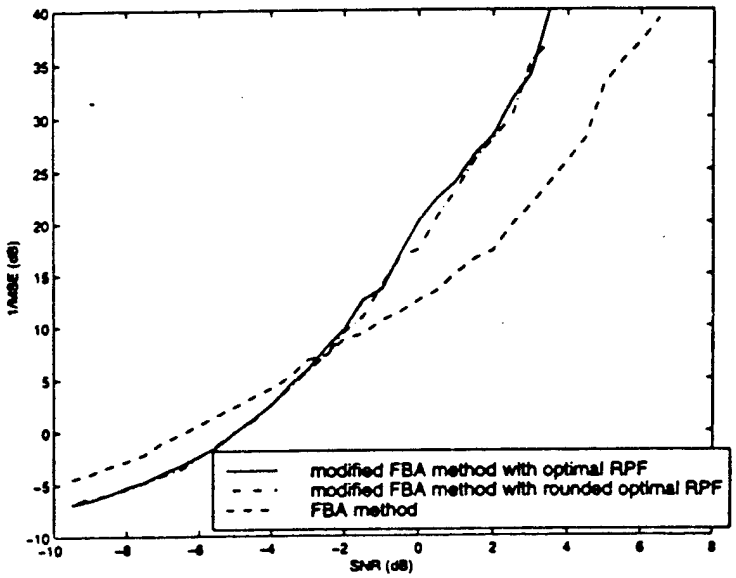

Fig 5 Comparison of reciprocal MSE of Doppler frequency estimations: FBA method and modified FBA method with optimal PRFs (or α_0) Solid line modified FBA method with optimal $\alpha_0 = 0.57$; dashdot line: modified FBA method with rounded optimal $\alpha_0 = 0.6$; dashed line: FBA method. $N(1) = 40$, $n_s = 12$, maximal ambiguity order $n_{\max} = 3$.

holds. In this case, we use the modified FBA method for the Doppler frequency detection. 20,000 Monte Carlo tests are implemented, i.e., $M = 20{,}000$ in (43). Three curves are plotted in Fig. 5 for the reciprocal MSE, 1/MSE, of the Doppler frequency estimations. The solid line is for the modified FBA method with the optimal $\alpha = 0.57$; the dashdot line is for the modified FBA method with the rounded α_0, 0.6; the dashed line is for the FBA method. A significant improvement of the MSE at the transition SNR band can be clearly seen.

As a remark, when $\alpha_0 = 1$, the FBA method and the modified FBA method both work. From our numerous numerical examples, these two methods have the same performance in this case.

When $n_{\max} = 12$ and $N(1) = -N(2) = 40 < 4(1 + 12) = 52$, i.e., the condition (19) or (11) for the accurate circular mean formula (12) does not hold. In this case, the FBA two step method does not work as shown in Fig. 6 and the modified FBA method should be used. The optimal parameter α_0 is $\hat{\alpha}_0 = 2.01$. 10,000 Monte Carlo tests are implemented, i.e., $M = 10{,}000$ in (43). Similar to Fig. 5, three curves are plotted in Fig. 6 for the reciprocal MSEs. The solid line is for the modified FBA method with the optimal $\alpha_0 = 2.01$. The dashdot line is for the modified FBA

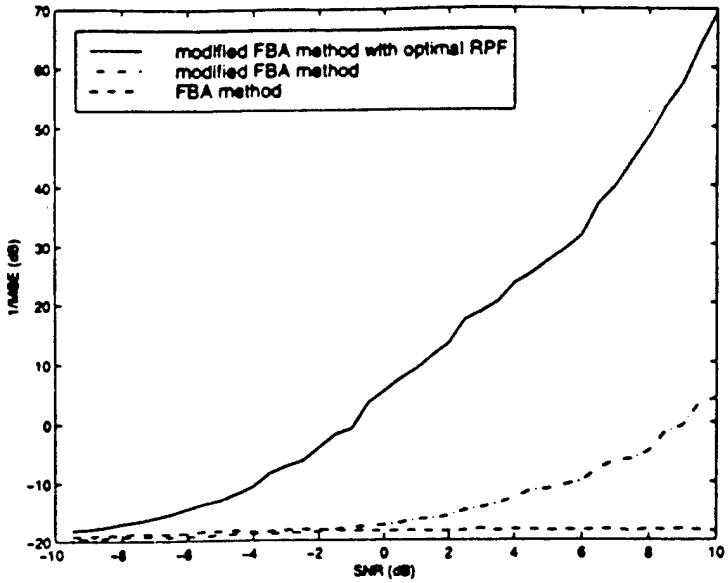

Fig. 6. Comparison of reciprocal MSE of Doppler frequency estimations: FBA method, modified FBA method, and modified FBA method with optimal PRFs (or α_0). Solid line: modified FBA method with optimal $\alpha_0 = 2.01$; dashdot line: modified FBA method; dashed line: FBA method. $N(1) = 40$, $n_s = 12$, maximal ambiguity order $n_{max} = 12$.

method with $\alpha_0 = 1$. The dashed line is for the FBA method. From Fig. 6, one can clearly see that in this case the FBA method fails, and the modified FBA method with the optimal α_0 outperforms the one with nonoptimal α_0.

V. CONCLUSION

In this paper, we studied the Ferrari–Bérenguer–Alengrin's two step Doppler frequency detection method, where the folded frequency is first estimated using the circular mean and the ambiguity order is then estimated using the quasi maximum likelihood criterion. The accuracy of the folded frequency depends on the use of the particular pairs of PRFs. When the folded frequency is not equal to the circular mean, we modified the FBA method with *a* finite possibilities of the folded frequency and the ambiguity order. More importantly, we studied and formulated the optimal PRFs in the modified FBA method in terms of minimizing the total sidelobe energy of the maximum likelihood function. Better performance of the modified FBA method over the FBA method was shown by numerical examples.

ACKNOWLEDGMENT

The author would like to thank the referees for their careful reading of this manuscript and one of the referees pointing out an error in its original version.

XIANG-GEN XIA
Dept. of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716
Email: (xxia@ee.udel.edu)

REFERENCES

[1] Ferrari, A., Bérenguer, C., and Alengrin, G. (1997) Doppler ambiguity resolution using multiple PRF. *IEEE Transactions on Aerospace and Electronic Systems*, **33** (July 1997), 738–751.

[2] Chang, C., and Curlander, J. (1992) Application of the multiple PRF technique to resolve Doppler centroid estimation ambiguity for spaceborne SAR. *IEEE Transactions on Geoscience and Remote Sensing*, **30** (Sept. 1992), 941–949.

[3] Ludloff, A., and Minker, M. (1985) Reliability of velocity measurement by MTD radar. *IEEE Transactions on Aerospace and Electronic Systems*, **AES-21** (July 1985), 522–528.

[4] Skolnik, M. (1981) *Introduction to Radar Systems*. New York: McGraw-Hill, 1981.

[5] Lovell, B., Kootsookos, P., and Williamson, R. (1991) The circular nature of discrete-time frequency estimates. In *Proceedings of the International Conference on Acoustics, Speech, Signal Processing* (1991), 3369–3372.

ELSEVIER Linear Algebra and its Applications 286 (1999) 19–35

LINEAR ALGEBRA AND ITS APPLICATIONS

Ambiguity resistant polynomial matrices

Guangcai Zhou [1], Xiang-Gen Xia *

Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716-3130, USA

Received 10 June 1997; accepted 19 June 1998

Submitted by L. Rodman

Abstract

An $N \times K$ $(N \geqslant K)$ ambiguity resistant (AR) matrix $G(z)$ is an irreducible polynomial matrix of size $N \times K$ over a field $\mathbf{F}$ such that the equation $EG(z) = G(z)V(z)$ with E an unknown constant matrix and $V(z)$ an unknown polynomial matrix has only the trivial solution $E = \alpha I_N$, $V(z) = \alpha I_K$, where $\alpha \in \mathbf{F}$. AR matrices have been introduced and applied in modern digital communications as error control codes defined over the complex field. In this paper we systematically study AR matrices over an infinite field $\mathbf{F}$. We discuss the classification of AR matrices, define their normal forms, find their simplest canonical forms, and characterize all $(K+1) \times K$ AR matrices that are the most interesting matrices in the applications.

AMS classification: 15A21; 15A24; 94A12; 94B10; 94B12

Keywords: Irreducible matrix; Ambiguity resistant matrix; Polynomial matrix; Error control coding

1. Background and introduction

An error control code defined over the complex field $\mathbb{C}$ maps each K input samples into N output samples, where N is usually greater than K so that the code is used to resist errors in a channel and the code is called $N \times K$ code. An

* Corresponding author. Tel.: 302 831-8038; fax: 302 831-4316; e-mail: xxia@ee.udel.edu.
[1] E-mail: gzhou@ee.udel.edu.

0024-3795/99/$ – see front matter
PII: S0024-3795(98)10127-1

$N \times K$ linear error control code is usually represented by an $N \times K$ polynomial matrix $G(z)$, where each entry of the matrix $G(z)$ is a polynomial of the variable z (or the delay variable z^{-1} in engineering) over complex field $\mathbb{C}$. Let $X(z)$ be a $K \times 1$ polynomial matrix (or vector) as an input signal. Then $Y(z) = G(z)X(z)$, is the $N \times 1$ polynomial matrix (or vector) of the code output signal, which is usually transmitted through a real world channel, wired or wireless. In a channel there are two common distortions. One is an additive random noise and the other is the so-called intersymbol interference (ISI). An additive random noise means that the received signal is $\tilde{Y}(z) = G(z)X(z) + \eta(z)$ instead of $Y(z)$, where $\eta(z)$ is the polynomial vector of the additive noise. Notice that the above additive noise only affects each individual sample of the received signal. The ISI is another type of distortion in a channel, which is usually due to a high speed transmission and cause distortions between received samples. Mathematically, the ISI is an $M \times N$ polynomial matirx $H(z)$ and the received signal is

$$\tilde{Y}(z) = H(z)Y(z) = H(z)G(z)X(z), \tag{1.1}$$

where $G(z)$ is an error control code.

Resistance to an additive random noise means that the input signal $X(z)$ can be restored from the received $\tilde{Y}(z)$ distorted by an additive noise $\eta(z)$ without knowing $\eta(z)$. To achieve this goal, the distance between codewords $Y(z) = G(z)X(z)$ after an error control code plays the most important role, see for example Ref. [1], which is beyond the scope of this paper. Similarly, resistance to the ISI means that the input signal $X(z)$ can be recovered from the received $\tilde{Y}(z)$ in Eq. (1.1) distorted by an ISI $H(z)$ without knowing $H(z)$. To achieve this goal, ambiguity resistant (AR) matrices have been introduced in Refs. [6,7], which are based on the irreducibility of polynomial matrices defined over the complex field $\mathbb{C}$. In this paper, a general infinite field $\mathbf{F}$ is considered. In what follows, $\mathbf{F}$ denotes an infinite field unless otherwise specified.

The definition of irreducible polynomial matrices over $\mathbf{F}$ induced from Ref. [3] is as follows.

Definition 1. An $N \times K$ $(N \geqslant K)$ polynomial matrix $G(z)$ over $\mathbf{F}$ is irreducible if and only if there is no $K \times K$ polynomial matrix $R(z)$ over $\mathbf{F}$ with non-constant determinant such that $G(z) = Q(z)R(z)$, where $Q(z)$ is an $N \times K$ polynomial matrix over $\mathbf{F}$.

The irreducibility can be characterized by the following lemma which offers an easy method to judge the irreducibility of a matrix when $\mathbf{F}$ is algebraically closed.

Lemma 1. *An $N \times K$ $(N \geqslant K)$ polynomial matrix $G(z)$ over an algebraically closed field $\mathbf{F}$ is irreducible if and only if it is full column rank (i.e., $\operatorname{rank}(G(z)) = K$) for any $z \in \mathbf{F}$.*

Remark 1. In Ref. [3] the definition of the irreducibility of a polynomial matrix over $\mathbb{C}$ is given by the necessary and sufficient condition in Lemma 1. However, when $\mathbf{F}$ is not algebraically closed, the irreducibility in Definition 1 is not equivalent to the necessary and sufficient condition in Lemma 1. As an example, polynomial matrix $(z^2+1)(z^2+2)I_N$ has full rank for any real $z \in \mathbb{R}$. It is, however, obviously reducible.

Definition 2. An $N \times K$ $(N \geqslant K)$ irreducible polynomial matrix $G(z)$ over $\mathbf{F}$ is called AR if and only if the following equation

$$EG(z) = G(z)V(z) \tag{1.2}$$

with E an unknown constant matrix and $V(z)$ an unknown polynomial matrix over $\mathbf{F}$ has only the trivial solution $E = \alpha I_N$, $V(z) = \alpha I_K$, where $\alpha \in \mathbf{F}$, I_N and I_K are $N \times N$ and $K \times K$ identity matrices, respectively.

It has been proved in Refs. [6,7] that, if the code $G(z)$ in Eq. (1.1) over the complex field $\mathbb{C}$ is AR, then the input signal $X(z)$ can be blindly recovered from the received signal $\tilde{Y}(z)$ in Eq. (1.1), where the knowledge of the ISI channel $H(z)$ is not necessary. Therefore, the resistance of the ISI can be achieved by choosing a code $G(z)$ to be AR.

Some necessary conditions for a code $G(z)$ over $\mathbb{C}$ to be AR are given in Ref. [6], for example, $G(z)$ is not a constant matrix, and $N > K$. Furthermore, it has been proved in Refs. [6,7] that the following $N \times (N-1)$ codes $G(z)$ are AR over $\mathbb{C}$

$$G(z) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ z^r & 1 & 0 & \cdots & 0 & 0 \\ 0 & z^r & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & z^r & 1 \\ 0 & 0 & 0 & \cdots & 0 & z^r \end{bmatrix}_{N\times(N-1)} \tag{1.3}$$

for any positive integer r. In Ref. [8], the following $N \times (N-1)$ polynomial matrices

$$G(z) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ F_1(z) & F_2(z) & F_3(z) & \cdots & F_{N-1}(z) \end{bmatrix}_{N\times(N-1)} \tag{1.4}$$

have been studied. The following necessary and sufficient condition for $G(z)$ in Eq. (1.4) to be AR is given in Ref. [8]:

$$\{1, F_1(z), F_2(z), \ldots, F_{N-1}(z)\}$$

is linearly independent over $\mathbb{C}$, which can be also seen in Theorem 4 in this paper for a general infinite field $\mathbf{F}$. The codes $G(z)$ in Eq. (1.4) are called *systematic codes*, which is analogous to the conventional error control codes defined over a finite field [2] for the encoding convenience.

In this paper, we systematically study AR matrices over $\mathbf{F}$. We provide canonical forms for all $N \times K$ AR matrices and characterize all $N \times (N-1)$ AR matrices. The characterization is easy to use. Since, in coding applications, K samples are expanded to N samples, when an $N \times K$ code is used. This expansion means that the bandwidth needs to be expanded in a transmission, which is usually expensive. Therefore, the smallest sample expansion in coding is usually desired. Clearly, the codes of size $N \times (N-1)$ provide the smallest bandwidth expansions, and therefore are the most interesting codes in applications. The characterization of all $N \times (N-1)$ AR matrices provides the opportunity to search the optimal one in resisting other distortions, such as the additive random noise as mentioned before. Some results have been obtained in Refs. [8,9] along this direction.

This paper is organized as follows. In Section 2 we present the canonical forms for $N \times K$ AR matrices. In Section 3, we provide the necessary and sufficient conditions for a polynomial matrix of size $N \times (N-1)$ to be AR in terms of its systematic or canonical form.

2. Classification of AR matrices and canonical form

Let $\mathbf{F}[z]$ denote the polynomial ring over an infinite field $\mathbf{F}$. Let $\mathbf{M}_{N\times K}(\mathbf{F}[z])$ denote the set of all $N \times K$ matrices with elements in $\mathbf{F}[z]$.

Definition 3. The transformation $T_{P,Q}$ of $\mathbf{M}_{N\times K}(\mathbf{F}[z])$ defined by

$$T_{P,Q}(A) = PAQ \quad \text{for all } A \in \mathbf{M}_{N\times K}(\mathbf{F}[z]),$$

where P and Q are $N \times N$ and $K \times K$ unimodular polynomial matrices (i.e., their determinants are non-zero constants), is called an equivalence transformation of $\mathbf{M}_{N\times K}(\mathbf{F}[z])$.

It is well known that the set of all equivalence transformations of $\mathbf{M}_{N\times K}(\mathbf{F}[z])$ is a group of transformations with $T_{I_M J_K}$ the identity transformation, and with the formulas $T_{P,Q}T_{R,S} = T_{PR,SQ}$ and $T_{P,Q}^{-1} = T_{P^{-1},Q^{-1}}$. This group induces an equivalence relation on $\mathbf{M}_{N\times K}(\mathbf{F}[z])$ and two matrices A and B are

said to be *equivalent over* $\mathbf{F}$ if there exists an equivalence transformation $T_{P,Q}$ such that $T_{P,Q}(A) = B$. Since polynomial ring $\mathbf{F}[z]$ is a principal ideal ring, we know that every matrix $A \in \mathbf{M}_{N\times K}(\mathbf{F}[z])$is equivalent to a diagonal matrix $D(z)$, which is known as the Smith form decomposition [4,5]. This general theory applies to AR matrices, but the equivalence relation defined above does not preserve the AR property. To do so, we define AR-equivalence transformations as follows.

Definition 4. An equivalence transformation $T_{P,Q}$ is called an AR-equivalence transformation if and only if P is a non-singular constant matrix and Q is a unimodular polynomial matrix.

From now on, in order to avoid confusion, A will represent a constant matrix and $A(z)$ will represent a polynomial matrix unless otherwise specified. We have the following result.

Theorem 1. *An AR-equivalence transformation preserves the AR property, i.e., an $N \times K$ polynomial matrix $G(z)$ is ambiguity resistant if and only if $PG(z)Q(z)$ is ambiguity resistant for any $N \times N$ invertible constant matrix P and any unimodular polynomial matrix $Q(z)$.*

Proof. Let $E_1PG(z)Q(z) = PG(z)Q(z)V_1(z)$. Then $P^{-1}E_1PG(z) = G(z)Q(z)V_1(z)Q^{-1}(z)$. Hence the ambiguity resistance of $G(z)$ implies that $Q(z)V_1(z)Q^{-1}(z) = \alpha I_K$ and $P^{-1}E_1P = \alpha I_N$, which implies $E_1 = \alpha I_N$, $V_1(z) = \alpha I_K$ for some non-zero constant $\alpha \in \mathbf{F}$. On the other hand, if $PG(z)Q(z)$ is AR, $EG(z) = G(z)V(z)$ means $PEP^{-1}[PG(z)Q(z)] = [PG(z)Q(z)]Q^{-1}(z)V(z)Q(z)$ which means $PEP^{-1} = \alpha I_N$ and $Q^{-1}(z)V(z)Q(z) = \alpha I_K$ for some non-zero $\alpha \in \mathbf{F}$. Hence $E = \alpha I_N$, $V(z) = \alpha I_K$ and $G(z)$ is AR. □

We can easily check that AR-equivalence transformations form a subgroup of equivalence transformations. They also induce an equivalence relation among AR matrices. We call $G_1(z)$ and $G_2(z)$ *AR-equivalent* if there is an AR-equivalence transformation $T_{P,Q}$ such that $T_{P,Q}G_1(z) = G_2(z)$.

For an irreducible $N \times K$ polynomial matrix $G(z)$, we can check that $G(z)$ is equivalent to matrix $[I_K, \mathbf{0}]^{\mathrm{T}}$ (if $N > K$) where A^{T} means the transpose of matrix A and $\mathbf{0}$ is the $K \times (N-K)$ matrix with 0 entries. We now want to seek a simple form of matrix $G(z)$ under AR-equivalence, which is useful for AR characterization. The following proposition was given in Ref. [6], where result (a) is useful later as a necessary condition on AR matrices, and result (b) makes us only need to consider the case $N > K$.

Proposition 1. *If an $N \times K$ ($N \geqslant K$) polynomial matrix $G(z)$ over $\mathbf{F}$ is AR, then*
(a) *$G(z)$ is not AR-equivalent to a matrix whose first column is $(1, 0, \ldots, 0)^{\mathrm{T}}$;*
(b) *$N > K$.*

Proof. (a) Suppose $G(z)$ is AR-equivalent to a matrix with first column $(1,0,\ldots,0)^{\mathrm{T}}$, then by simple equivalence it follows that $G(z)$ is AR-equivalent to a matrix of the form

$$G_1(z) = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & H(z) \end{bmatrix},$$

where $H(z)$ is an $(N-1)\times(K-1)$ polynomial matrix. Setting

$$E = \begin{bmatrix} 2 & \mathbf{0} \\ \mathbf{0} & I_{N-1} \end{bmatrix}, \quad \text{and} \quad V(z) = \begin{bmatrix} 2 & \mathbf{0} \\ \mathbf{0} & I_{K-1} \end{bmatrix},$$

we see that $EG_1(z) = G_1(z)V(z)$ and $V(z) \neq \alpha I_K$ for any non-zero constant $\alpha \in \mathbf{F}$. In other words, $G(z)$ is not AR.

(b) If $N = K$, then the irreducibility of $G(z)$ means $G(z)$ is unimodular. So for any E, $V(z) = G^{-1}(z)EG(z)$ satisfies $EG(z) = G(z)V(z)$. So $G(z)$ is not AR. □

Lemma 2. *Any polynomial matrix* $A(z) \in \mathbf{M}_{N\times K}(\mathbf{F}[z])$ *with* $\operatorname{rank} = K$ *is AR-equivalent to*

$$\begin{bmatrix} g_{11}(z) & 0 & 0 & \ldots & 0 \\ g_{21}(z) & g_{22}(z) & 0 & \ldots & 0 \\ g_{31}(z) & g_{32}(z) & g_{33}(z) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ g_{K1}(z) & g_{K2}(z) & g_{K3}(z) & \ldots & g_{KK}(z) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ g_{N1}(z) & g_{N2}(z) & g_{N3}(z) & \ldots & g_{NK}(z) \end{bmatrix}, \tag{2.1}$$

where $\deg(g_{11}(z)) \leqslant \deg(g_{22}(z)) \leqslant \cdots \leqslant \deg(g_{KK}(z))$. *Furthermore,* $\deg(g_{ij}(z)) < \deg(g_{ii}(z))$ *for any* $j < i$.

Proof. Let $A(z)$ be an $N \times K$ matrix with entries $a_{ij}(z)$. Let $d_i(z) = \mathrm{GCD}(a_{i1}(z),\ldots,a_{iK}(z))$. By row permutation only we may assume that $d_i(z) \neq 0$ and $\deg d_i$ is non-decreasing with i for $i = 1,\ldots,K$. Now $A(z)$ is AR-equivalent to (by only column transforms)

$$\begin{bmatrix} d_1(z) & 0 & \ldots & 0 \\ b_{21}(z) & b_{22}(z) & \ldots & b_{2K}(z) \\ \ldots & \ldots & \ldots & \ldots \\ b_{N1}(z) & b_{N2}(z) & \ldots & b_{NK}(z) \end{bmatrix}.$$

Furthermore, $\deg[\mathrm{GCD}(b_{i2}(z),\ldots,b_{iK}(z))] \geqslant \deg[\mathrm{GCD}(b_{i1}(z),\ldots,b_{iK}(z))] \geqslant \deg d_i \geqslant \deg d_1(z)$ for $i = 2,\ldots,N$. Similarly we can deal with the submatrix

$$B(z) = \begin{bmatrix} b_{22}(z) & \dots & b_{2K}(z) \\ \dots & \dots & \dots \\ b_{N2}(z) & \dots & b_{NK}(z) \end{bmatrix}$$

with $\operatorname{rank}(B) = K - 1$. By induction the lemma is proved. □

Remark 2. Form (2.1) has a direct relationship with row-Hermite forms and the above lemma can also be proved by using Theorem 6.3-2 in Ref. [3]. From Ref. [3] the row-Hermite form of a matrix $A(z)$ is equal to $A(z)Q(z)$ where $Q(z)$ is a unimodular $K \times K$ matrix. By row permutations, it is guaranteed that the diagonal elements $g_{jj}(z)$ are non-zero and $\deg g_{ii} \leqslant \deg g_{jj}$ if $1 \leqslant i \leqslant j \leqslant K$. Using column operations again the polynomial matrix can be reduced to the form in Lemma 2.

Lemma 3. *For L polynomials $f_1(z) \neq 0, f_2(z), \dots, f_L(z)$ over $\mathbf{F}$, if* $\deg(\mathrm{GCD}(cf_1 + f_2, f_3, \dots, f_L)) \geqslant \deg f_1$ *for any constant $c \in \mathbf{F}$, then $f_1 | f_2, f_1 | f_3, \dots, f_1 | f_L$.*

Proof. We first prove the case $L = 3$. It is obvious if f_1 is a constant. Now suppose $\deg f_1 \geqslant 1$ and $d_c(z) \equiv \mathrm{GCD}(cf_1 + f_2, f_3)$. Then $\deg d_c \geqslant 1$. Let $d = \mathrm{GCD}(f_1, f_2, f_3)$. Then $f_1 = dg_1, f_2 = dg_2, f_3 = dg_3$ and $\mathrm{GCD}(g_1, g_2, g_3) = 1$, $\deg(\mathrm{GCD}(cg_1 + g_2, g_3)) \geqslant \deg g_1$. But based on the fact that if $\mathrm{GCD}(g_1, g_2, g_3) = 1$ over an infinite field $\mathbf{F}$ then there exists $c \in \mathbf{F}$ such that $\mathrm{GCD}(cg_1 + g_2, g_3) = 1$. Hence the above two cases mean $\mathrm{GCD}(cg_1 + g_2, g_3) = 1$. Therefore we have $\mathrm{GCD}(cf_1 + f_2, f_3) = d \times \mathrm{GCD}(cg_1 + g_2, g_3) = d$. Now $\deg d \geqslant \deg f_1$ and $d | f_1$ imply $d(z) = cf_1(z)$ for some non-zero constant c. Hence $f_1 | f_2$, $f_1 | f_3$. For general L we know $\deg(\mathrm{GCD}(cf_1 + f_2, f_3, \dots, f_L)) = \deg(\mathrm{GCD}(cf_1 + f_2, \mathrm{GCD}(f_3, \dots, f_L)))$. By the above proof, $f_1 | f_2, f_1 | \mathrm{GCD}(f_3, \dots, f_L)$. Hence $f_1 | f_2, f_1 | f_3, \dots, f_1 | f_L$. □

Lemma 4. *If $G(z) = (g_{ij}(z))$ is a non-zero matrix in $\mathbf{M}_{N \times K}(\mathbf{F}[z])$ of the form* (2.1) *and if $g_{11}(z) \neq 0$ is an element of $G(z)$ with $m = \deg(g_{11}) \leqslant \deg(g_{ij})$ for any $g_{ij}(z)$, then either $g_{11}(z)$ divides all $g_{ij}(z)$, or else there exists an AR-equivalence transform T such that $T(G(z)) = H(z)$ has the form* (2.1) *and $h_{11}(z) \neq 0$ is of degree less than m.*

Proof. Suppose $g_{11}(z)$ does not divide every element of $G(z)$. By Lemma 3, there exists a constant $c \in \mathbf{F}$ and $i, 2 \leqslant i \leqslant N$ such that $\deg(\mathrm{GCD}(cg_{11} + g_{i1}, g_{i2}, \dots, g_{ii}, \dots g_{iK}) < \deg g_{11} = m$. This means that $G(z)$ is AR-equivalent to a matrix with i-row $(cg_{11} + g_{i1}, g_{i2}, \dots, g_{ii}, \dots, g_{iK})$. Now Lemma 2 guarantees that $G(z)$ is AR-equivalent to $H(z)$ of form (2.1) with $\deg h_{11} \leqslant \deg(\mathrm{GCD}(cg_{11} + g_{i1}, g_{i2}, \dots, g_{ii}, \dots, g_{iK}) < m$. □

Combining Lemmas 2 and 4 we can obtain the following result.

Theorem 2. *Any non-zero matrix* $A(z) \in \mathbf{M}_{N\times K}(\mathbf{F}[z])$ *with* $\operatorname{rank} = k$ *is AR-equivalent to a matrix of the following form*

$$\begin{bmatrix} g_{11}(z) & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ g_{21}(z) & g_{22}(z) & 0 & \dots & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ g_{k1}(z) & g_{k2}(z) & g_{k3}(z) & \dots & g_{kk}(z) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ g_{N1}(z) & g_{N2}(z) & g_{N3}(z) & \dots & g_{Nk}(z) & 0 & \dots & 0 \end{bmatrix}$$

with $g_{ii}|g_{(i+1)(i+1)}, g_{ii}|g_{ji}$ for any $i = 1, 2, \dots, k-1$ and $j \geqslant i$.

Proof. Obviously, $A(z)$ is AR-equivalent to a matrix of form $[B\ \mathbf{0}]$ where B is an $N \times k$ matrix with $\operatorname{rank}(B) = k$. By Lemma 2, we have that any non-zero matrix is AR-equivalent to a matrix as above such that $g_{11}(z)$ has the minimum degree. If $g_{ii}(z)$ divides all $g_{kl}(z)$ for any $k, l \geqslant i$, Theorem 2 is proved. If $g_{ii}(z)$ does not divide some $g_{kl}(z)$ for some $k, l \geqslant i$, we then consider the submatrix

$$\begin{bmatrix} g_{ii}(z) & 0 & 0 & \dots & 0 \\ g_{(i+1)i}(z) & g_{(i+1)(i+1)}(z) & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ g_{ki}(z) & g_{k(i+1)}(z) & g_{k(i+2)}(z) & \dots & g_{kk}(z) \\ \dots & \dots & \dots & \dots & \dots \\ g_{Ni}(z) & g_{N(i+1)}(z) & g_{N(i+2)}(z) & \dots & g_{Nk}(z) \end{bmatrix}.$$

Therefore, by Lemma 4, under AR-equivalence we have that $g_{ii}(z)$ divides all $g_{kl}(z)$ for any $k, l \geqslant i$. □

By the above theorem, for irreducible matrices, we have the following result.

Theorem 3. *Any irreducible matrix in* $\mathbf{M}_{N\times K}(\mathbf{F}[z])$ *is AR-equivalent to a matrix of the following form*

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ g_{K1}(z) & g_{K2}(z) & g_{K3}(z) & \dots & g_{K(K-1)}(z) & g_{KK}(z) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ g_{N1}(z) & g_{N2}(z) & g_{N3}(z) & \dots & g_{N(K-1)}(z) & g_{NK}(z) \end{bmatrix} \tag{2.2}$$

with $\mathrm{GCD}(g_{KK}, g_{(K+1)K}, \ldots, g_{NK}) = 1$, $\deg g_{iK} < \deg g_{jK}$ *for* $K \leqslant i < j \leqslant N$. *Furthermore,* $g_{kl}(z)$ *can be either* 0 *or a non-constant polynomial* (*i.e.,* $\deg g_{kl} \geqslant 1$) *for* $K \leqslant k \leqslant N$ *and* $1 \leqslant l \leqslant K-1$, *and* $g_{N1}(z) = \cdots = g_{N(L-1)}(z) = 0$, $1 \leqslant \deg g_{NL} < \deg g_{N(L+1)} < \cdots < \deg g_{NK}$ *for some* L *where* $1 \leqslant L < K$.

Proof. By Theorem 2, if $g_{ii}(z)$ is not a non-zero constant for $1 \leqslant i \leqslant K-1$, then $g_{ji}(z) = g_{ii}(z)h_{ji}(z)$ for $i < j \leqslant N$. For example, assume $i = 1$. Then

$$G(z) = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ h_{21}(z) & g_{22}(z) & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ h_{K1}(z) & g_{K2}(z) & g_{K3}(z) & \ldots & g_{KK}(z) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ h_{N1}(z) & g_{N2}(z) & g_{N3}(z) & \ldots & g_{NK}(z) \end{bmatrix} \begin{bmatrix} g_{11}(z) & \mathbf{0} \\ \mathbf{0} & I_{K-1} \end{bmatrix}$$

which contradicts with the irreducibility of $G(z)$ because the leftmost matrix is not unimodular. Similar arguments can be used to prove that $\mathrm{GCD}(g_{KK}, g_{(K+1)K}, \cdots, g_{NK}) = 1$. When $g_{kl}(z)$ is a non-zero constant for some k,l with $K \leqslant k \leqslant N$ and $1 \leqslant l \leqslant K-1$, it can be reduced to zero by implementing a constant elementary row operation, i.e., $g_{kl}(z)$ can be reduced to zero by an AR-equivalence transformation. □

Remark 3. The result in Theorem 3 is the simplest form we can have, which cannot be improved further. For example, we can directly check that

$$\begin{bmatrix} 1 & 0 \\ z & z^2 \\ z^2 & z^3+1 \end{bmatrix}$$

is an irreducible matrix, we cannot simplify it further under AR-equivalence transformations. This polynomial matrix is actually an AR matrix from Theorem 5 of the next section.

Definition 5. If $A(z)$ is AR-equivalent to $G(z)$ of the form (2.2), then $G(z)$ is called the canonical form of $A(z)$. If $g_{KK} = 1$, as indicated in Section 1, we call $G(z)$ the systematic form of $A(z)$.

By the above results, we can easily classify irreducible matrices as well as AR matrices. So, to study the AR property of a polynomial matrix, we only need to study its canonical form or systematic form.

3. $(K+1) \times K$ AR-matrices

In the above sections we have discussed the classification of AR matrices and it was shown that every $N \times K$ AR matrix is of form (2.2). In this section, we present the sufficient and necessary conditions for a $(K+1) \times K$ matrix to be AR. These conditions can be used in the design of error control codes in applications.

We first see the simplest form, i.e., the systematic form as follows (also see Ref. [8]).

Theorem 4. *If $G(z)$ has systematic form, i.e.,*

$$G(z) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ g_{(K+1)1}(z) & g_{(K+1)2}(z) & g_{(K+1)3}(z) & \dots & g_{(K+1)K}(z) \end{bmatrix}_{(K+1)\times K}$$

then $G(z)$ is AR if and only if $\{1, g_{(K+1)1}(z), g_{(K+1)2}(z), g_{(K+1)3}(z), \dots, g_{(K+1)K}(z)\}$ are linearly independent over **F**.

Proof. Let $N = K+1$. We first prove the necessity. If $\{1, g_{N1}(z), g_{N2}(z), g_{N3}(z), \dots, g_{NK}(z)\}$ are linearly dependent, then there exists $k \in \{1, \dots, K\}$ such that

$$g_{NK}(z) = c + \sum_{i=1, i\neq k}^{K} c_i g_{Ni}(z).$$

Hence, there exists an AR-equivalence transform that transforms $G(z)$ into a matrix with its first column as $(1, 0, \dots, 0)^{\mathrm{T}}$. Proposition 1 means $G(z)$ is not AR. Thus

$$\{g_{N1}(z), g_{N2}(z), g_{N3}(z), \dots, g_{NK}(z)\}$$

are linearly independent.

We now prove the sufficiency. Under AR-equivalence, we may assume $1 \leqslant \deg g_{N1} < \deg g_{N2} < \cdots < \deg g_{NK}$. By

$$\begin{bmatrix} e_{11} & \dots & e_{1N} \\ \dots & \dots & \dots \\ e_{N1} & \dots & e_{NN} \end{bmatrix} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ g_{N1}(z) & g_{N2}(z) & \dots & g_{NK}(z) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ g_{N1}(z) & g_{N2}(z) & \dots & g_{NK}(z) \end{bmatrix} \begin{bmatrix} v_{11}(z) & \dots & v_{1K}(z) \\ \dots & \dots & \dots \\ v_{K1}(z) & \dots & v_{KK}(z) \end{bmatrix}$$

we obtain

$$e_{ij} + e_{iN}g_{Nj}(z) = v_{ij}(z) \quad \text{for } i,j = 1,2,\dots,K, \tag{3.1}$$

$$e_{Nj} + e_{NN}g_{Nj}(z) = \sum_{k=1}^{K} v_{kj}(z)g_{Nk}(z) \quad \text{for } j = 1,2,\dots,K. \tag{3.2}$$

First, from Eq. (3.1) and Eq. (3.2) we obtain

$$e_{Nj} + e_{NN}g_{Nj}(z) = \sum_{k=1}^{K-1}(e_{kj} + e_{kN}g_{Nj}(z))g_{Nk}(z) + (e_{Kj} + e_{KN}g_{Nj}(z))g_{NK}(z).$$

Taking $j = K$ we have

$$e_{NK} + e_{NN}g_{NK}(z) = \sum_{k=1}^{K-1}(e_{kK} + e_{kN}g_{NK}(z))g_{Nk}(z) + (e_{KK} + e_{KN}g_{NK}(z))g_{NK}(z).$$

Comparing the highest coefficients of the two polynomials we have $e_{KK} = e_{NN}$ and $e_{kN} = 0$ for any $k = 1,\dots,K$. Hence $v_{ij}(z) = e_{ij}$ is in fact a constant for any $i,j = 1,\dots,K$. Since $1 \leqslant \deg(g_{N1}) < \deg(g_{N2}) < \deg(g_{NK})$, $1, g_{N1},\dots,g_{NK}$ are linearly independent. By Eq. (3.2) again we have

$$e_{Nj} + e_{NN}g_{Nj}(z) = \sum_{k=1}^{K} v_{kj}g_{Nk}(z) \quad \text{for } j = 1,2,\dots,K.$$

Hence $v_{ij} = e_{ij} = 0$ except possibly $v_{jj} = e_{jj} = e_{NN}$ for $j = 1,2,\dots,K$. This is exactly what we need. □

We now consider the general canonical form (2.2) $G(z)$ with $N = K + 1$. We have the following necessary and sufficient conditions for all possible $(K + 1) \times K$ AR matrices.

Theorem 5. *Let $G(z)$ have the cononical form (2.2) with $N=K+1$.*

(a) *If $1 \leqslant \deg g_{N1} < \deg g_{N2} < \cdots < \deg g_{NK}$, then $G(z)$ is AR if and only if $\mathrm{GCD}(g_{KK}, g_{NK}) = 1$. In this case, irreducibility and ambiguity resistance are the same.*

(b) *If $g_{N1}(z) = \cdots = g_{N(L-1)}(z) = 0, 1 \leqslant \deg g_{NL} < \cdots < \deg g_{NK}$, then $G(z)$ is AR if and only if $\mathrm{GCD}(g_{KK}, g_{NK}) = 1$, $\{1, g_{K1}, g_{K2}, \ldots, g_{K(L-1)}, g_{NL}, \ldots, g_{N(K-1)}\}$ are linearly independent over $\mathbf{F}$, and $W_1 \cap W_2 = \{0\}$, where*

$$W_1 = \mathrm{span}\{g_{NK}, g_{NK}g_{K1}, \ldots, g_{NK}g_{K(K-1)}\},$$

$$W_2 = \mathrm{span}\{g_{KK}, g_{KK}g_{K1}, \ldots, g_{KK}g_{K(L-1)}, g_{KK}g_{NL}, \ldots, g_{KK}g_{N(K-1)}\},$$

where span means the set of all linear combinations with constant coefficients.

Proof. (a) This is a special case of (b): the case of $L=1$. If $\mathrm{GCD}(g_{NK}, g_{KK}) = 1$ and $1 \leqslant \deg g_{N1} < \cdots < \deg g_{NK}$, we can easily check that $\mathrm{span}\{g_{NK}, g_{NK}g_{K1}, \ldots, g_{NK}g_{K(K-1)}\} \cap \mathrm{span}\{g_{KK}, g_{KK}g_{N1}, \ldots, g_{KK}g_{N(K-1)}\} = \{0\}$. So we only need to prove (b).

(b) $EG(z) = G(z)V(z)$ we get the following equations:

$$e_{ij} + e_{iK}g_{Kj}(z) + e_{iN}g_{Nj}(z) = v_{ij}(z), \quad 1 \leqslant i, j \leqslant K-1, \tag{3.3}$$

$$e_{Kj} + e_{KK}g_{Kj}(z) + e_{KN}g_{Nj}(z) = \sum_{m=1}^{K} g_{Km}(z)v_{mj}(z), \quad 1 \leqslant j \leqslant K-1, \tag{3.4}$$

$$e_{Nj} + e_{NK}g_{Kj}(z) + e_{NN}g_{Nj}(z) = \sum_{m=1}^{K} g_{Nm}(z)v_{mj}(z), \quad 1 \leqslant j \leqslant K-1, \tag{3.5}$$

$$e_{iK}g_{KK}(z) + e_{iN}g_{NK}(z) = v_{iK}(z), \quad 1 \leqslant i \leqslant K-1, \tag{3.6}$$

$$e_{KK}g_{KK}(z) + e_{KN}g_{NK}(z) = \sum_{m=1}^{K} g_{Km}(z)v_{mK}(z), \tag{3.7}$$

$$e_{NK}g_{KK}(z) + e_{NN}g_{NK}(z) = \sum_{m=1}^{K} g_{Nm}(z)v_{mK}(z). \tag{3.8}$$

Substituting Eq. (3.6) to Eq. (3.8) we obtain

$$e_{NK}g_{KK}(z) + e_{NN}g_{NK}(z) = \sum_{m=1}^{K-1} g_{Nm}(z)(e_{mK}g_{KK}(z) + e_{mN}g_{NK}(z)) + v_{KK}(z)g_{NK}(z),$$

i.e.,

$$\left(e_{NN} - v_{KK}(z) - \sum_{m=1}^{K-1} e_{mN} g_{Nm}(z)\right) g_{NK}(z) = \left(\sum_{m=1}^{K-1} e_{mK} g_{Nm}(z) - e_{NK}\right) g_{KK}(z).$$

So $\mathrm{GCD}(g_{NK}, g_{KK}) = 1$ implies

$$g_{NK}(z) \left| \left(\sum_{m=1}^{K-1} e_{mK} g_{Nm}(z) - e_{NK}\right) \right. .$$

Hence $1 \leqslant \deg g_{NL} < \cdots < \deg g_{NK}$ implies $e_{NK} = 0$ and $e_{iK} = 0$ for $i = L, \ldots, K-1$ and

$$v_{KK}(z) = e_{NN} - \sum_{m=1}^{K-1} e_{mN} g_{Nm}(z). \tag{3.9}$$

Plugging Eqs. (3.6) and (3.9) into Eq. (3.7) we get

$$\left(e_{KN} - \sum_{m=1}^{K-1} e_{mN} g_{Km}(z)\right) g_{NK}(z)$$
$$= \left(e_{NN} - e_{KK} + \sum_{m=1}^{K-1} (e_{mK} g_{Km} - e_{mN} g_{Nm}(z))\right) g_{KK}(z),$$

or

$$\left(e_{KN} - \sum_{m=1}^{K-1} e_{mN} g_{Km}(z)\right) g_{NK}(z)$$
$$= \left(e_{NN} - e_{KK} + \sum_{m=1}^{L-1} e_{mK} g_{Km} - \sum_{m=L}^{K-1} e_{mN} g_{Nm}(z)\right) g_{KK}(z). \tag{3.10}$$

Now $W_1 \cap W_2 = \{0\}$ and the linear independence of $\{1, g_{K1}, \ldots, g_{K(L-1)}, g_{NL}, \ldots, g_{N(K-1)}\}$ mean $e_{KK} = e_{NN}$, $e_{iK} = 0$ for $i = 1, \ldots, L-1$, $e_{mN} = 0$ for $m = L, \ldots, K-1$. So

$$e_{KN} - \sum_{m=1}^{K-1} e_{mN} g_{Km}(z) = e_{KN} - \sum_{m=1}^{L-1} e_{mN} g_{Km}(z) = 0$$

implies $e_{KN} = 0$, $e_{mN} = 0$ for $m = 1, \ldots, L-1$ by the linear independence of $\{1, g_{K1}, \ldots, g_{K(L-1)}\}$. Hence we obtain $e_{iN} = 0$ for $i = 1, 2, \ldots, K$, $e_{KN} = 0$, $e_{KK} = e_{NN}$, $e_{iK} = 0$ for $i = 1, 2, \ldots, K-1$. Then Eq. (3.3) becomes

$$v_{ij}(z) = e_{ij}, \quad i, j = 1, 2, \ldots, K-1, \tag{3.11}$$

Eq. (3.4) becomes

$$e_{Kj} + e_{KK}g_{Kj}(z) = \sum_{m=1}^{K} g_{Km}(z)v_{mj}(z), \quad 1 \leqslant j \leqslant K-1, \tag{3.12}$$

Eq. (3.5) becomes

$$e_{Nj} + e_{NN}g_{Nj}(z) = \sum_{m=1}^{K} g_{Nm}(z)v_{mj}(z) = \sum_{m=L}^{K} g_{Nm}(z)v_{mj}(z),$$
$$1 \leqslant j \leqslant K-1. \tag{3.13}$$

Plugging Eq. (3.11) into Eq. (3.13) we get $e_{NN} = e_{jj} = v_{jj}, v_{ij}(z) = e_{ij} = 0$ for $i = L, \ldots, K$, $j = 1, 2, \ldots, K-1$, $i \neq j$. In this case Eq. (3.12) becomes

$$e_{Kj} + e_{KK}g_{Kj}(z) = \sum_{m=1}^{L-1} g_{Km}(z)e_{mj} + \sum_{m=L}^{K} g_{Km}(z)e_{mj}.$$

This means that $e_{ij} = 0$ if $i = 1, \ldots, L-1$, $j = 1, 2, \ldots, K-1$, $i \neq j$. This proves that $E = e_{11}I_{K+1}, V(z) = e_{11}I_K$.

We now prove the necessity. If $G(z)$ is AR, it is obvious that we require $\mathrm{GCD}(g_{NK}, g_{KK}) = 1$ and $\{1, g_{K1}, \ldots, g_{K(L-1)}\}$ are linearly independent. Now if $W_1 \cap W_2 \neq \{0\}$, Eq. (3.10) implies that we can find non-trivial solution, i.e., there exists $e_{iK} \neq 0$ for some $1 \leqslant i \leqslant L-1$. Hence we conclude that $V(z) \neq \alpha I_K$. This contradicts with the AR property of $G(z)$. □

Remark 4. In Theorem 5, if $g_{KK}(z) = 1$ and $g_{Kj}(z) = 0$ for $1 \leqslant j \leqslant K-1$, it is exactly Theorem 4.

By Theorem 5, we can also see that if the field **F** is the complex field $\mathbb{C}$, and the degree of the polynomial matrix of size $(K+1) \times K$ to be bounded by some integer M, then the set of $(K+1) \times K$ polynomial matrices that are not AR has measure 0 in the finite dimensional linear space consisting of all $(K+1) \times K$ polynomial matrices whose degrees are bounded by M. This means that a randomly generated polynomial matrix is AR with probability 1. Hence we have the following result which confirms the conjecture made in Ref. [10].

Corollary 1. *A randomly generated $(K+1) \times K$ polynomial matrix over the complex field $\mathbb{C}$ is almost surely AR.*

The following corollary gives an intuitive construction of a family of AR matrices.

Corollary 2. *If* $g_{N1}(z) = \cdots = g_{N(L-1)}(z) = 0$, $1 \leqslant \deg g_{NL} < \cdots < \deg g_{NK}$, $\mathrm{GCD}(g_{KK}, g_{NK}) = 1$, $\{1, g_{K1}, g_{K2}, \ldots, g_{K(L-1)}, g_{NL}, \ldots, g_{N(K-1)}\}$ *are linearly independent over* **F** *and if* $\deg g_{NK} > \deg g_{Kj}$ *for* $1 \leqslant j \leqslant L-1$, *then* $G(z)$ *is AR.*

Proof. Let

$$W_1 = \mathrm{span}\{g_{NK}, g_{NK}g_{K1}, \ldots, g_{NK}g_{K(K-1)}\},$$

$$W_2 = \mathrm{span}\{g_{KK}, g_{KK}g_{K1}, \ldots, g_{KK}g_{K(L-1)}, g_{KK}g_{NL}, \ldots, g_{KK}g_{N(K-1)}\}.$$

We only need to prove $W_1 \cap W_2 = \{0\}$. Now let

$$\sum_{j=1}^{K-1} \alpha_j g_{NK} g_{Kj} = \sum_{j=1}^{L-1} \beta_j g_{KK} g_{Kj} + \sum_{j=L}^{K-1} \beta_j g_{KK} g_{Nj}$$

i.e.,

$$g_{NK}\left(\sum_{j=1}^{K-1} \alpha_j g_{Kj}\right) = g_{KK}\left(\sum_{j=1}^{L-1} \beta_j g_{Kj} + \sum_{j=L}^{K-1} \beta_j g_{Nj}\right).$$

By $\mathrm{GCD}(g_{KK}, g_{NK}) = 1$ we get

$$g_{NK} \left| \left(\sum_{j=1}^{L-1} \beta_j g_{Kj} + \sum_{j=L}^{K-1} \beta_j g_{Nj}\right).\right.$$

So $\deg g_{NK} > \deg g_{Kj}$ and $\deg g_{NL} < \cdots < \deg g_{NK}$ induce

$$\sum_{j=1}^{L-1} \beta_j g_{Kj} + \sum_{j=L}^{K-1} \beta_j g_{Nj} = 0$$

and hence $W_1 \cap W_2 = \{0\}$. □

It is natural to ask the following question: if $G(z)$ of size $N \times K$ $(N > K)$ is AR, $H(z)$ is an $M \times K$ polynomial matrix, is polynomial matrix

$$\begin{bmatrix} G(z) \\ H(z) \end{bmatrix}$$

AR? The following example provides a negative answer to this question.

Example 1. Let

$$G(z) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z & z^2 \end{bmatrix}.$$

By Theorem 4 we see that $G(z)$ is AR. Let

$$\tilde{G}(z) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z & z^2 \\ 0 & z \end{bmatrix}.$$

We can easily check that

$$\begin{bmatrix} 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z & z^2 \\ 0 & z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ z & z^2 \\ 0 & z \end{bmatrix} \begin{bmatrix} 1-z & -z^2 \\ 1 & 1+z \end{bmatrix}$$

Hence $\tilde{G}(z)$ is not AR.

However, we have the following property [6].

Proposition 2. *If an $M \times K$ polynomial matrix $A(z)$ is AR-equivalent to*

$$\begin{bmatrix} G(z) \\ \mathbf{0} \end{bmatrix}$$

and $G(z)$ is AR, then $A(z)$ is AR.

Proof. We only need to prove that if $G(z)$ is AR of size $N \times K$, then

$$\begin{bmatrix} G(z) \\ \mathbf{0} \end{bmatrix}$$

is AR. By equation

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \begin{bmatrix} G(z) \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} G(z) \\ \mathbf{0} \end{bmatrix} V(z)$$

we get $E_{11}G(z) = G(z)V(z)$, so $G(z)$ is AR concludes $V(z) = \alpha I_K$ for some nonzero constant α. □

In Section 3 we completely characterized $(K+1) \times K$ AR matrix. However, the sufficient conditions for general $N \times K$ polynomial matrices to be AR are not yet clear. Another interesting question is, if $G(z)$ is an $N \times K$ $(N > K+1)$ AR matrix, can we always find an ambiguity resistant $(K+1) \times K$ submatrix $H(z)$ among the AR-equivalence class of $G(z)$?

Finally, as pointed out by one of the referees, some of the results in this paper also apply to finite fields. For instance, let us consider Lemma 3. Since the degrees of f_i for $1 \leqslant i \leqslant L$ are bounded, using a simple non-topological counting argument, Lemma 3 is also true for a sufficiently large finite field. Since Lemma 3 plays a main role in the proof of Theorems 2 and 3, we believe

that if the degrees of polynomial matrices are bounded and the matrix size is fixed, Theorems 2 and 3 is also true for a sufficiently large finite field. Nevertheless, we think that the results in Theorems 2–5 may not hold for general finite fields when there is no restriction on polynomial matrices.

Acknowledgements

The authors would like to thank the associate editor and one of the anonymous referees for their insightful comments and suggestions that have improved the clarity of this manuscript. In particular, they wish to thank one of the referees for the suggestion of describing the results in this paper over a general infinite field rather than the complex field $\mathbb{C}$ only. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, the National Science Foundation CAREER program under Grant MIP-9703377, and the University of Delaware Research Foundation.

References

[1] C.E. Shannon, A mathematical theory of communications, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[2] S. Lin, D.J. Costello, Jr., Error Control Coding: Fundamentals and Applications, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[3] T. Kailath, Linear Systems, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[4] I. Gohberg, P. Lancaster, L. Rodman, Matrix Polynomials, Academic Press, New York, 1982.
[5] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[6] H. Liu, X.-G. Xia, Precoding techniques for undersampled multi-receiver communication systems, Technical Report #97-3-1, Department of Electrical Engineering, University of Delaware, 1997.
[7] X.-G. Xia, H. Liu, Polynomial ambiguity resistant precoders: theory and applications in ISI/multipath cancelation, Technical Report #97-5-1, Department of Electrical Engineering, University of Delaware, 1997.
[8] X.-G. Xia, G. Zhou, On optimal ambiguity resistant precoders in ISI/multipath cancelation, preprint, Technical Report #97-5-2, Department of Electrical Engineering, University of Delaware, 1997.
[9] X.-G. Xia, Modulated coding and least square decoding via coded modulation and Viterbi decoding, Technical Report #97-6-2, Department of Electrical and Computer Engineering, University of Delaware, 1997.
[10] H. Liu, Private communications, 1997.

which completes the derivation with $\Upsilon(\dot{L}_{22}) = \kappa_{22}\{\|I - L_{21}L_{11}^{-1}L_{12}D^{-1}\|_1 + \|P_{22}\|_1\}$.

Equation (14) indicates that the error made in approximating the 22 block in L is proportional to $\Delta(\dot{L}_{22})$; however, Υ is a function of L_{22} through α and the 1-norms of P_{22} and $I - L_{11}^{-1}L_{21}^{-1}L_{12}D^{-1}$. To partially examine the behavior of Υ, we argue that decreasing $\Delta(\dot{L}_{22})$ in a natural manner decreases this quantity as well. Assume that Θ is fixed and that Δ is decreased by decresing the value of the κ parameter in the prior model, specifically for the coefficients in the 22 block. Now, it is not hard to show that $D^{-1} \xrightarrow{\kappa=0} 0$ so that with $B = L_{21}L_{11}^{-1}L_{12}$

$$\begin{aligned}\|I - BD^{-1}\|_1 &= \|I - BD^{-1} + D^{-1} - D^{-1}\|_1 \\ &\le \|I - D^{-1}\|_1 + \|D^{-1} - BD^{-1}\|_1 \\ &\le \|I - D^{-1}\|_1 \\ &\quad + \|D^{-1}\|_1\|I - B\|_1 \xrightarrow{\kappa=0} \|I\|_1 = 1.\end{aligned}$$

Hence, asymptotically, $\|I - BD^{-1}\|_1$ is independent of $\dot{L}_{22}$. Referring to (8b), it is not difficult to show that as κ decreases, $P_{22} \to \kappa^2 P_{0,22}$, where $P_{0,22}$ is the 22 block of the appropriately permuted form of P_0. Therefore, as $\kappa \to 0$, both α and $\|P_{22}\|_1$ go to 0. Thus, we conclude that decresing $\Delta(\dot{L}_{22})$ by varying the degree of regularization will cause $\Upsilon \to 0$.

REFERENCES

[1] B. Alpert, G. Beylkin, R. Coifman, and V. Rokhlin, "Wavelets for the fast solution of second-kind integral equations," *SIAM J. Sci. Comput.*, vol. 14, no. 1, pp. 159–184, 1993.

[2] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 2024–2036, Dec. 1989.

[3] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[4] E. L. Miller, "The application of multiscale and statistical techniques to the solution of inverse problems," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, Aug. 1994.

[5] E. L. Miller and A. S. Willsky, "A multiscale approach to sensor fusion and the solution of linear inverse problems," *Appl. Comput. Harmon. Anal.*, vol. 2, pp. 127–147, 1995.

[6] ——, "Multiscale, statistically-based inversion scheme for the linearized inverse scattering problem," *IEEE Trans. Geosci. Remote Sensing*, vol. 34, pp. 346–357, Mar. 1996.

[7] ——, "Wavelet-based methods for the nonlinear inverse scattering problem using the extended Born approximation," *Radio Sci.*, vol. 31, no. 1, pp. 51–65, Jan./Feb. 1996.

Efficient Implementation of Arbitrary-Length Cosine-Modulated Filter Bank

Xiqi Gao, Zhenya He, and Xiang-Gen Xia

Abstract— **The fast implementation of arbitrary-length cosine-modulated filter bank is investigated. By using the linear phase property of the prototype filter, a more efficient implementation structure is obtained for the filter bank. In the new implementation, 2 × 2 lossless lattices are used instead of 2 × 1 ones in the traditional implementation with the number reduced by half.**

I. INTRODUCTION

The cosine-modulated filter bank (CMFB) has received much interest in recent years [1]–[6]. It has two remarkable features: easy design and fast implementation. While the design of CMFB's has been addressed by many researchers, we deal with the implementation of paraunitary CMFB's in this correspondence. Typically, the polyphase component matrix of a paraunitary CMFB can be expressed as the product of a modulation part and a polyphase part in terms of the polyphase components of the prototype filter. Based on such an expression, the CMFB can be implemented through two-channel lossless lattices and fast discrete cosine/sine transform (DCT/DST) algorithms (see, for example, [1] and [4]). Two-channel lattices are often used for an M-channel CMFB. Notice that only half the number of the lattices are required in the implementation of Malvar's CMFB, which is called extended lapped transform (ELT) [2]. The motivation of this correspondence is to generalize the above Malvar's result to other paraunitary CMFB's. The arbitrary-length CMFB developed by Nguyen and Koilpillai in [3] is considered in this correspondence.

This correspondence is organized as follows. In Section II, we first review the arbitrary-length CMFB briefly. Then, we show that the four filters in two related pairs of power complementary polyphase components of the prototype filter form a 2 × 2 paraunitary system due to the prototype filter symmetry. In Section III, a new expression of the polyphase component matrix of the CMFB is developed. Based on it, a more efficient implementation structure is obtained by using the 2 × 2 lossless lattices instead of the 2 × 1 ones in the traditional implementation. The implementation complexity of the CMFB is discussed in Section IV.

Notations: Capital and lower case letters are used to denote the transfer functions and the impulse responses of filters, respectively. Bold letters indicate vectors and matrices. The functions $\lceil x \rceil$ and $\lfloor x \rfloor$ round the value of x to the nearest integers toward infinity and minus infinity, respectively. $\boldsymbol{C}_N^{\mathrm{III}}$ and $\boldsymbol{C}_N^{\mathrm{IV}}$ stand for the standard DCT matrices as defined in [8]. $\boldsymbol{0}$ stands for matrix whose entries are all zeros. $\boldsymbol{I}_N$ and $\boldsymbol{J}_N$ are the $N \times N$ identity and reverse identity matrices, respectively.

Manuscript received August 7, 1997; revised August 15, 1998. This work was supported in part by Natural Science Foundation of China. The associate editor coordinating the review of this paper and approving it for publication was Dr. Sergios Theodoridis.

X. Gao and Z. He are with the Department of Radio Engineering, Southeast University, Nanjing, China.

X.-G. Xia is with the Department of Electrical Engineering, University of Delaware, Newark, DE 19716 USA.

Publisher Item Identifier S 1053-587X(99)02163-7.

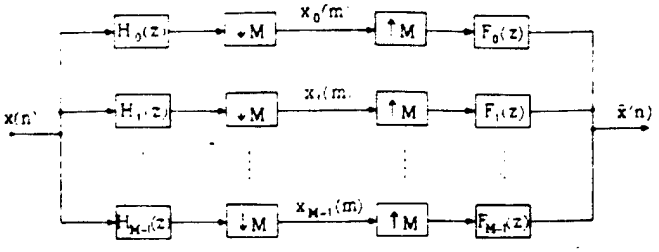

Fig. 1. M-channel maximally decimated filter bank.

II. The Arbitrary-Length CMFB

A. A Review of The Arbitrary-Length CMFB

A typical M-channel maximally decimated filter bank is shown in Fig. 1, where $H_k(z)$ and $F_k(z)(0 \le k \le M-1)$ are the transfer functions of the analysis and synthesis filters, respectively. At the analysis side, the input signal $x(n)$ is decomposed into M subband signals through the bank of analysis filters followed by M-fold decimators. At the synthesis side, M subband signals are passed through M-fold interpolators and recombined into the reconstructed signal $\hat{x}(n)$ by using the bank of synthesis filters.

Let $h(n)$ denote the impulse response of a linear-phase low-pass prototype filter with length $N = 2m_0M + m_1$, where m_0 and m_1 are integers and $0 \le m_1 \le 2M-1$. The M-channel arbitrary-length CMFB is defined as [3]

$$h_k(n) = 2h(n)\cos\left(\frac{\pi(k+0.5)}{M}\left(n-\frac{N-1}{2}\right)+(-1)^k\frac{\pi}{4}\right) \tag{1a}$$

$$f_k(n) = 2h(n)\cos\left(\frac{\pi(k+0.5)}{M}\left(n-\frac{N-1}{2}\right)-(-1)^k\frac{\pi}{4}\right) \tag{1b}$$

where $h_k(n)$ and $f_k(n), 0 \le k \le M-1, 0 \le n \le N-1$, are the impulse responses of the kth analysis and synthesis filters, respectively. The CMFB is exactly the one investigated in [1] with the length extended to arbitrary integer value.

Suppose that the impulse response of the lowpass prototype filter is symmetry; then, $f_k(n)$ is the time-reversed and shifted version of $h_k(n)$. This relation means that the CMFB is paraunitary if and only if it has perfect reconstruction property [9]. Let $G_k(z), k = 0, 1, \cdots, 2M-1$ denote the type-1 polyphase components of the prototype filter [9]. Due to the symmetry property of $h(n)$, $G_k(z)$ satisfies

$$\tilde{G}_k(z) = \begin{cases} z^{m_1}G_{m_1-1-k}(z), & k \le m_1-1 \\ z^{m_1-1}G_{2M+m_1-1-k}(z), & k \ge m_1 \end{cases} \tag{2}$$

where $\tilde{G}_k(z) = G_k(z^{-1})$. It has been shown in [3] that the necessary and sufficient condition on the polyphase components for perfect reconstruction is

$$\tilde{G}_k(z)G_k(z) + \tilde{G}_{M+k}(z)G_{M+k}(z) = \frac{1}{2M}, \qquad 0 \le k \le M-1. \tag{3}$$

This means that appropriate pairs of the polyphase filters are power complementary. Depending on the lengths of the two filters $G_k(z)$ and $G_{M+k}(z)$ and the relationship between them, four classes of power complementary pairs can be distinguished in the general case for arbitrary length prototype filters. The condition given by (3) can be satisfied by the four different modes, as discussed in [3]. In *modes* a and c, the two filters have the same length. If they are related by (2), they are under *mode* c; otherwise, they are under *mode* a. In *mode* c, both of the two filters must be delays. In *mode* b and d, $G_k(z)$ and $G_{M+k}(z)$ have different lengths. If they are related to themselves by (2), they are under *mode* d; otherwise, they are under *mode* b. In *mode* d, one of the two filters must be a delay, and all coefficients of the other one must be zeros.

B. Lattice Structure for a Power Complementary Pair and Its Related One

The power complementary filter pair $G_k(z)$ and $G_{M+k}(z)$ satisfying (3) can be completely factored as the two-channel lossless lattice

$$\sqrt{2M}\begin{bmatrix} G_k(z) \\ G_{M+k}(z) \end{bmatrix} = \boldsymbol{R}_{k,m}\Lambda(z)\boldsymbol{R}_{k,m-1}\Lambda(z)\cdots\boldsymbol{R}_{k,1}\Lambda(z)\begin{bmatrix} c_{k,0} \\ s_{k,0} \end{bmatrix} \tag{4}$$

where

$$\boldsymbol{R}_{k,l} = \begin{bmatrix} c_{k,l} & s_{k,l} \\ s_{k,l} & -c_{k,l} \end{bmatrix}, \quad \Lambda(z) = \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix}$$
$$c_{k,l} = \cos\theta_{k,l}, \quad s_{k,l} = \sin\theta_{k,l}, \qquad l = 0, 1, 2, \cdots, m$$

and m depends on the lengths of the two filters. For the case $N = 2mM$, all the polyphase components have the same length, and there is no restriction on any angle parameter $\theta_{k,l}$. For the general case when the prototype filter has arbitrary length, there are different constraints on the angles of the lossless lattices corresponding to the four modes (see [3] for details).

For each power complementary filter pair, we can find a related one due to the prototype filter symmetry of (2). The four filters in the two pairs define a 2×2 system. We define the following three types of 2×2 systems in terms of different m_1 and k:

$$\boldsymbol{L}_k^{(1)}(z) \triangleq \sqrt{2M}\begin{bmatrix} G_k(z) & G_{M+m_1-1-k}(z) \\ -z^{-1}G_{M-k}(z) & G_{m_1-1-k}(z) \end{bmatrix}$$
$$m_1 \le M, \quad k \le m_1-1 \quad \text{or}$$
$$m_1 > M, \quad m_1 - M \le k \le M-1 \tag{5a}$$

$$\boldsymbol{L}_k^{(2)}(z) \triangleq \sqrt{2M}\begin{bmatrix} G_k(z) & G_{M+m_1-1-k}(z) \\ G_{M+k}(z) & -G_{2M+m_1-1-k}(z) \end{bmatrix}$$
$$m_1 \le M, \quad m_1 \le k \le M-1 \tag{5b}$$

$$\boldsymbol{L}_k^{(3)}(z) \triangleq \sqrt{2M}\begin{bmatrix} G_k(z) & G_{m_1-M-1-k}(z) \\ G_{M+k}(z) & -G_{m_1-1-k}(z) \end{bmatrix}$$
$$m_1 > M, \quad k \le m_1-M-1. \tag{5c}$$

$\boldsymbol{L}_k^{(1)}$ is for the polyphase filters under *mode* b and *mode* d. $\boldsymbol{L}_k^{(2)}$ and $\boldsymbol{L}_k^{(3)}$ are for the polyphase filters under *mode* a and *mode* c. It is easy to show that these systems are paraunitary and can be expressed as the following 2×2 lattices:

$$\boldsymbol{L}_k^{(1)}(z) = \begin{cases} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}\boldsymbol{R}_{k,m_0-1}\Lambda(z)\cdots\boldsymbol{R}_{k,1}\Lambda(z) \\ \quad \boldsymbol{R}_{k,0}\begin{bmatrix} z^{-1} & 0 \\ 0 & (-1)^{m_0} \end{bmatrix}, \qquad \theta_{k,m_0} = \dfrac{\pi}{2} \\ \Lambda(-z)\boldsymbol{R}_{k,m_0}\Lambda(z)\cdot\boldsymbol{R}_{k,2}\Lambda(z)\boldsymbol{R}_{k,1} \\ \quad \begin{bmatrix} 1 & 0 \\ 0 & (-1)^{(m_0+1)} \end{bmatrix}, \qquad \theta_{k,0} = 0. \end{cases} \tag{6a}$$

$$\boldsymbol{L}_k^{(2)}(z) = \boldsymbol{R}_{k,m_0-1}\Lambda(z)\boldsymbol{R}_{k,m_0-2}\Lambda(z) \cdots \boldsymbol{R}_{k,1}\Lambda(z)\boldsymbol{R}_{k,0}\begin{bmatrix} 1 & 0 \\ 0 & (-1)^{m_0} \end{bmatrix} \tag{6b}$$

$$\boldsymbol{L}_k^{(3)}(z) = \boldsymbol{R}_{k,m_0}\Lambda(z)\boldsymbol{R}_{k,m_0-1}\Lambda(z) \cdots \boldsymbol{R}_{k,1}\Lambda(z)\boldsymbol{R}_{k,0}\begin{bmatrix} 1 & 0 \\ 0 & (-1)^{m_0+1} \end{bmatrix}. \tag{6c}$$

In *mode c* and *mode d*, the corresponding lattices are trivial.

In practice, a two-channel lossless lattice can be implemented by using the two-multiplier structure for each section [2], [9]. A 2×1 lossless lattice with m free angle parameters and others set to be zero or $\pi/2$ can be implemented by using $2m$ multipliers and $2(m-1)$ adders. For the corresponding 2×2 system, both the two numbers are $2m+1$. For $\boldsymbol{L}_k^{(1)}$ and $\boldsymbol{L}_k^{(2)}$, there are m_0 free angle parameters, and hence, $2m_0+1$ multipliers and $2m_0+1$ adders are required. For $\boldsymbol{L}_k^{(3)}$, there are m_0+1 free angle parameters, and hence, $2m_0+3$ multipliers and $2m_0+3$ adders are required.

III. Fast Implementation of the Arbitrary-Length CMFB

Now, we consider the implementation of the CMFB. Considering the relationship between the synthesis bank and the analysis bank of the paraunitary filter bank, we only deal with the latter. The polyphase component matrix of the analysis bank can be expressed as [3]

$$\boldsymbol{E}(z) = \dot{\boldsymbol{C}}\begin{bmatrix} \boldsymbol{g}_0(-z^2) \\ z^{-1}\boldsymbol{g}_1(-z^2) \end{bmatrix} \tag{7}$$

where

$$[\dot{\boldsymbol{C}}]_{k,l} = 2\cos\left(\frac{\pi(k+0.5)}{M}\left(l - \frac{N-1}{2}\right) + (-1)^k\frac{\pi}{4}\right)$$

$$\boldsymbol{g}_0(z) = \operatorname{diag}(G_0(z) \quad G_1(z) \quad \cdots \quad G_{M-1}(z)), \text{ and}$$

$$\boldsymbol{g}_1(z) = \operatorname{diag}(G_M(z) \quad G_{M+1}(z) \quad \cdots \quad G_{2M-1}(z)).$$

Based on (7) and the power complementary condition in (3) for perfect reconstruction, the filter bank can be implemented through a parallel bank of 2×1 lossless lattices cascaded by the modulation matrix. The number of the 2×1 lattices is equal to the number of subchannels M. The modulation part can be implemented by fast DCT algorithm. Such an implementation structure has been widely used in the CMFB's with $N = 2mM$ [1], [4]. The linear-phase property of the prototype filter is not exploited to reduce the complexity.

Let $m_1 + M - 1 = 2l_0 - l_1$, where l_0 is an integer, and $l_1 \in \{0, 1\}$. If l_1 is equal to zero, one of m_1 and M is odd, and the other even. If l_1 is equal to one, both m_1 and M are odd or even. By using the properties of cosine function, the modulation matrix can be expressed as

$$\dot{\boldsymbol{C}} = \sqrt{2M}\boldsymbol{D}\boldsymbol{C}[\boldsymbol{A}\Lambda_0 - \boldsymbol{B}\Lambda_1 \quad \boldsymbol{A}\Lambda_1 + \boldsymbol{B}\Lambda_0] \tag{8}$$

where $\boldsymbol{D}$ is an $M \times M$ diagonal matrix with the kth diagonal component $d_k = (-1)^{k/2}$, and

$$\boldsymbol{C} = \begin{cases} \boldsymbol{C}_M^{III}, & l_1 = 0 \\ \boldsymbol{C}_M^{IV}, & l_1 = 1 \end{cases} \quad \boldsymbol{A} = \begin{cases} \begin{bmatrix} \sqrt{2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{M-1} \end{bmatrix}, & l_1 = 0 \\ \boldsymbol{I}_M, & l_1 = 1 \end{cases}$$

$$\boldsymbol{B} = \begin{cases} -\begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}_{M-1} \end{bmatrix}, & l_1 = 0 \\ -\boldsymbol{J}_M, & l_1 = 1 \end{cases}$$

$$\Lambda_0 = \begin{cases} \begin{bmatrix} 0 & \boldsymbol{I}_{M-l_0} \\ \boldsymbol{0} & 0 \end{bmatrix}, & l_0 \le M \\ -\begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{I}_{l_0-M} & \boldsymbol{0} \end{bmatrix}, & l_0 > M \end{cases}$$

$$\Lambda_1 = \begin{cases} \begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{I}_{l_0} & \boldsymbol{0} \end{bmatrix}, & l_0 \le M \\ \begin{bmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{2M-l_0} \end{bmatrix}, & l_0 > M. \end{cases}$$

Substituting (8) into (7), we obtain the following expression of $\boldsymbol{E}(z)$:

$$\boldsymbol{E}(z) = \boldsymbol{D}\boldsymbol{C}\boldsymbol{G}(z) \tag{9}$$

where

$$\boldsymbol{G}(z) = \sqrt{2M}[(\boldsymbol{A}\Lambda_0 - \boldsymbol{B}\Lambda_1)\boldsymbol{g}_0(-z^2) + z^{-1}(\boldsymbol{A}\Lambda_1 + \boldsymbol{B}\Lambda_0)\boldsymbol{g}_1(-z^2)]. \tag{10}$$

Based on this new expression, the analysis bank can be implemented more efficiently. The implementation structure is shown in Fig. 2. The diagonal matrix $\boldsymbol{D}$ only changes the signs of the output subband signals. The matrix $\boldsymbol{C}$ is the type III DCT and type IV DCT for l_1 to be zero and one, respectively. It can be shown that the $M \times M$ matrix $\boldsymbol{G}(z)$ can be implemented through a parallel bank of 2×2 lossless lattices that are related to $\boldsymbol{L}_k^{(1)}$ and some delays. To give the explicit formula of $\boldsymbol{G}(z)$ and see this clearly, eight cases can be distinguished in terms of l_0, l_1 and m_1, as shown in Table I. In Table I, we give the numbers of the 2×2 lattices used in the implementation of $\boldsymbol{G}(z)$ for the eight different cases. Here, we consider Case 2 as an example. In this case, $1 \le m_1 \le M-1$. $\boldsymbol{G}(z)$ takes the form of (11), shown at the bottom of the page, where G_k stands for $\sqrt{2M}G_k(-z^2)$. $\boldsymbol{G}(z)$ can be implemented in parallel through the following:

i) a delay $2\sqrt{M}G_{l_0}(-z^2)$ and another possible delay $\sqrt{2M}[G_{(m_1-1)/2}(-z^2) + z^{-1}G_{M+(m_1-1)/2}(-z^2)]$ when M is even;

$$\boldsymbol{G}(z) = \sqrt{2M}\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 00 & \sqrt{2}G_{l_0} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & G_{l_0-1} & 0 & G_{l_0+1} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \cdot & 0 & 0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 0 & G_{m_1} & 0 & 070 & 0 & 0 & G_{M-1} & \\
z^{-1}G_M & 0 & 0 & G_{m_1-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \cdot & \cdot & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
G_0 & 0 & 0 & z^{-1}G_{M+m_1-1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & z^{-1}G_{M+m_1} & 0 & 0 & 0 & 0 & 0 & -z^{-1}G_{2M-1} \\
0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 & \cdot & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & z^{-1}G_{M+l_0-1} & 0 & -z^{-1}G_{M+l_0+1} & 0 & 0
\end{bmatrix} \tag{11}$$

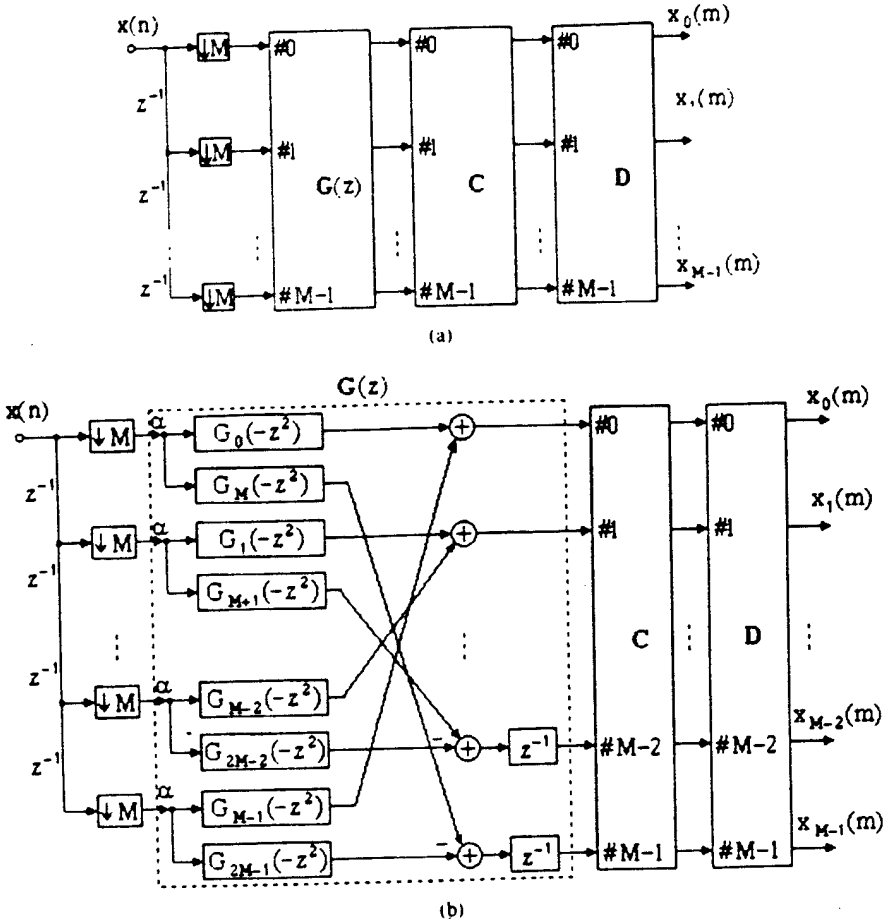

Fig. 2. New implementation structure of the M-channel CMFB. (a) For general case with arbitrary length. (b) For special case with $N = 2mM$, which covers Case 1 and Case 5. Here, $G(z)$ can be implemented in parallel through a bank of 2×2 lossless lattices in addition to some delays. C is the standard DCT (type III or type IV), and D is a diagonal matrix and only affects the signs of the outputs. In (b), α is equal to $\sqrt{2M}$.

ii) a set of 2×2 parauinary systems

$$L_k(z) = \sqrt{2M}\begin{bmatrix} G_k(-z^2) & G_{M+m_1-1-k}(-z^2) \\ z^{-1}G_{M+k}(-z^2) & -z^{-1}G_{2M+m_1-1-k}(-z^2) \end{bmatrix} = \Lambda(z)L_k^{(2)}(-z^2) \tag{12}$$

where $m_1 \le k \le (M + m_1 - 3)/2$;

iii) another set of 2×2 parauintary systems [see (13) at the bottom of the page];

where $0 \le k \le \lfloor m_1/2 \rfloor - 1$. Two types of the 2×2 parauintary systems defined in Section II are used in the parallel implementation. The total numbers of the 2×2 parauintary systems are $(M - m_1 - 1)/2$ and $\lfloor m_1/2 \rfloor$, respectively.

IV. Implementation Complexity

In the previous section, we have shown that the CMFB can be implemented through a parallel bank of 2×2 lossless lattices and some delays followed by the standard DCT. For the DCT's, fast algorithms are available [7], [8]. All the lattices are related to $L_k^{(i)}, i = 1, 2, 3$ without additional multipliers and adders required for the implementation. From Table I, it is obvious that the total number is less than or equal to $M/2$ for each case. The implementation cost of the CMFB is that of about $M/2$ 2×2 lattices plus one M-point DCT matrix working at an M-fold decimated rate.

In the traditional implementation structure of an M-channel CMFB, the number of the 2×1 lattices is M. Ignoring the trivial lattices, which are under *mode c* and *mode d*, the number is exactly twice as that in the new implementation structure. Since only one additional multiplier and three additional adders are required to implement the corresponding 2×2 lattice of a 2×1 one, the complexity of implementing a set of 2×2 lattices is lower than that of doubled 2×1 ones. When the section numbers are large, the complexities of the two type lattices are approached, and hence, the cost can be saved nearly one half to implement a set of 2×2 lattices instead of doubled 2×1 lattices.

As an example, we consider Case 5, with $M = 2^m$ to see the efficiency of the new implementation structure. In this case,

$$L_k(z) = \sqrt{2M}\begin{bmatrix} z^{-1}G_{M+k}(-z^2) & G_{m_1-1-k}(-z^2) \\ G_k(-z^2) & z^{-1}G_{M+m_1-1-k}(-z^2) \end{bmatrix} = \begin{bmatrix} 0 & z \\ 1 & 0 \end{bmatrix} L_k^{(1)}(-z^2)\Lambda(z)$$

$$= \begin{cases} \Lambda(-z)R_{k,m_0-1}\Lambda(-z^2)R_{k,m_0-2}\Lambda(-z^2)\cdots R_{k,1}\Lambda(-z^2)R_{k,0}\begin{bmatrix} z^{-1} & 0 \\ 0 & (-1)^{m_0+1} \end{bmatrix}, & \theta_{k,m_0} = \dfrac{\pi}{2} \\ \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} R_{k,m_0}\Lambda(-z^2)R_{k,m_0-1}\Lambda(-z^2)\cdots R_{k,2}\Lambda(-z^2)R_{k,1}\begin{bmatrix} 1 & 0 \\ 0 & (-1)^{m_0+1}z^{-1} \end{bmatrix}, & \theta_{k,0} = 0. \end{cases} \tag{13}$$

TABLE I
NUMBERS OF LATTICES IN THE NEW IMPLEMENTATION STRUCTURE

Case	Number of lattices		
	$\mathbf{L}_k^{(1)}$	$\mathbf{L}_k^{(2)}$	$\mathbf{L}_k^{(3)}$
1). $l_1=0$, $l_0<M$ and $m_1=0$	0	$\lfloor M/2 \rfloor$	0
2). $l_1=0$, $l_0<M$ and $m_1\neq 0$	$\lfloor m_1/2 \rfloor$	$\lfloor (M-m_1)/2 \rfloor$	0
3). $l_1=0$, $l_0=M$	$\lfloor (M-1)/2 \rfloor$	0	0
4). $l_1=0$, $l_0>M$	$M-\lfloor m_1/2 \rfloor$	0	$\lfloor (m_1-M-1)/2 \rfloor$
5). $l_1=1$, $l_0<M$ and $m_1=0$	0	$\lfloor M/2 \rfloor$	0
6). $l_1=1$, $l_0<M$ and $m_1\neq 0$	$\lfloor m_1/2 \rfloor$	$\lfloor (M-m_1)/2 \rfloor$	0
7). $l_1=1$, $l_0=M$	$\lfloor (M-1)/2 \rfloor$	0	0
8). $l_1=1$, $l_0>M$	$M-\lfloor m_1/2 \rfloor$	0	$\lfloor (m_1-M-1)/2 \rfloor$

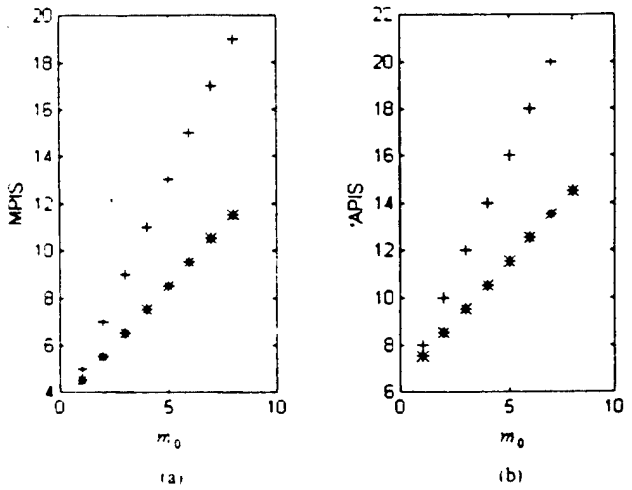

Fig. 3. Computational complexity in case 5 with $M = 16$. (a) Number of multiplications per input sample. (b) Number of additions per input sample. Here, + and * represent the complexities of the traditional and the new implementation structures, respectively.

m_1 is equal to zero. $M/2$ lattices $\boldsymbol{L}_k^{(2)}(z)$, and type IV DCT are used. By using the fast algorithm presented in [8], $(M/2)\log_2 M + M$ multiplications and $(3M/2)\log_2 M$ additions are required to compute the M-point DCT-IV. The total cost of implementing the CMFB is $M(2m_0 + 3 + \log_2 M)/2$ multiplications and $M(2m_0 + 1 + 3\log_2 M)/2$ additions per M input samples. It is the same cost required by Malvar's ELT with the same length [2]. In the traditional implementation, $M(4m_0 + 2 + \log_2 M)/2$ multiplications and $M(4m_0 + 3\log_2 M)/2$ additions are required for M input samples [1], [9]. In Fig. 3, we plot the average numbers of the multiplications and additions per input sample (MPIS and APIS) *versus* m_0 with $M = 16$. It can be seen that by using the new implementation structure, the saving of operations becomes more significant as m_0 increases.

V. CONCLUSION

A more efficient implementation structure for a class of paraunitary CMFB with arbitrary length has been developed in this correspondence. The linear-phase property of the prototype filter is exploited to reduce the implementation cost. The new implementation structure uses 2×2 lossless lattices instead of 2×1 ones with the total number of lattices reduced by half. The implementation costs are significantly saved, especially for the CMFB with a large ratio of the length to the number of channels.

REFERENCES

[1] R. D. Koilpillai and P. P. Vaidyanathan, "Cosine-modulated FIR filter banks satisfying perfect reconstruction," *IEEE Trans. Signal Processing*, vol. 40, pp. 770–783, Apr. 1992.

[2] H. S. Malvar, "Extended lapped transforms: properties, applications and fast algorithms," *IEEE Trans. Signal Processing*, vol. 40, pp. 2703–2714, Nov. 1992.

[3] T. Q. Nguyen and R. D. Koilpillai, "Theory and design of arbitrary-length cosine-modulated filter banks and wavelets, satisfying perfect reconstruction," *IEEE Trans. Signal Processing*, vol. 44, pp. 473–486, Mar. 1996.

[4] R. A. Gopinath and C. S. Burrus, "Theory of modulated filter banks and modulated wavelet tight frames," *Proc. IEEE ICASSP*, Minneapolis, MN, 1993, vol. III, pp. 169–172.

[5] G. D. T. Schuller and M. J. Smith, "New framework for modulated perfect reconstruction filter banks," *IEEE Trans. Signal Processing*, vol. 44, pp. 1941–1954, Aug. 1996.

[6] H. Xu, W. S. Lu, and A. Antoniou, "Efficient iterative design method for cosine-modulated QMF banks," *IEEE Trans. Signal Processing*, vol. 44, pp. 1657–1667, June 1996.

[7] Z. D. Wang, "Fast Algorithms for the discrete W transform and for the discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 803–817, Aug. 1984.

[8] ——, "On computing the discrete Fourier and cosine transforms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1341–1344, Oct. 1985.

[9] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

A Quantitative Analysis of SNR in the Short-Time Fourier Transform Domain for Multicomponent Signals

Xiang-Gen Xia*

August 12, 1999

Abstract

A quantitative analysis is given for the signal-to-noise ratio (SNR) in the short-time Fourier transform domain for multicomponent signals in additive white noise. It is shown that the SNR is increased on the order of $O(N/K)$ where K is the number of components of a signal, N/T is the sampling rate, and T is the window size. The SNR increase rate is optimal for given K. For this result, the SNR definition is generalized, which is suitable for signals not only in the time domain but also in other domains. This theory is illustrated by one numerical example.

1 Introduction

Time-frequency analysis [11-12] has become an important technique in analyzing wideband/nonstationary signals in various applications including inverse synthetic aperture radar (ISAR) imaging [1], biomedical signal analysis [2-3], speech signal analysis [4], and FM radio communications [5]. One of the most important features of this technique is that it usually increases the signal-to-noise ratio (SNR) in the joint time-frequency (TF) domain. This is particularly advantageous for signals which are difficult to detect in the time or frequency domain alone. The reason for this important feature can be stated as follows. A joint TF transform usually spreads noise from one dimension (the time or frequency) into two dimensions (the joint time and frequency) while it usually concentrates a signal in localized regions in the TF plane. A number of research results on the estimation of time-varying frequencies have appeared, such as [5-7] with Wigner-Ville distributions. However, to the author's best knowledge, there does not exist a quantitative analysis for the SNR increase for any joint time-frequency transform, which is certainly an important issue in practical applications in signal detection by using thresholding.

In the conventional SNR definition, the mean power is taken over the whole domain of a signal. If the signal is stationary in this domain, this definition works fine. However, if the signal is not stationary in this domain, such as a single tone signal in the frequency domain, this definition is no longer suitable. In this correspondence, we first generalize the SNR definition so that it is not only suitable for signals in the time domain but also in other domains, such as the frequency domain and the joint time-frequency

*Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716. Email: xxia@ee.udel.edu. Phone: (302)831-8038. Fax: (302)831-4316. This work was partially supported by an initiative grant from the Department of Electrical and Computer Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

domain. We then present a quantitative analysis of the SNR increase rate in the joint time-frequency domain for the short-time Fourier transform with rectangular windows, where multicomponent signals in additive white noise are considered. The main result can be stated as follows. Let K be the number of monocomponents in a signal, T be the window size for the short-time Fourier transform, and N/T be the sampling rate. N point discrete Fourier transform is performed in each window. Then, the SNR in the joint time-frequency domain is increased on the order of $O(N/K)$, when the window size T is small enough.

This correspondence is organized as follows. In Section 2, we formulate a proper definition for SNR in different domains. In Section 3, we present the proposed quantitative approach to analyze the SNR increase rate in the joint time-frequency domain. A numerical example is presented in Section 4 to illustrate the proposed approach.

2 Signal-to-Noise Ratio in Different Domains

The conventional signal-to-noise ratio (SNR) is defined as the ratio of the mean power of the signal over the mean power of the noise, where the mean is taken over the whole time domain. It is formulated as follows. Let $y[n]$ be a distorted signal:

$$y[n] = x[n] + \eta[n], \quad 0 \le n \le N-1, \tag{2.1}$$

where $x[n]$ is a signal and $\eta[n]$ is an additive white noise with variance σ^2. The SNR is defined as:

$$\text{SNR} = \frac{\sum_{n=0}^{N-1} |x[n]|^2}{N\sigma^2} \tag{2.2}$$

This SNR is used quite often in describing the noise level relative to the signal, and in distinguishing the signal from noise in stationary environments. When the SNR is too low, in general it is impossible to distinguish the signal $x[n]$ from $y[n]$. However, for some special kinds of signals $x[n]$, such as narrow band signals, it is possible to detect the signal in the Fourier transform domain even when the SNR is of negative dB. An example is shown in Fig. 1, where the SNR= -11dB and the signal x is a single tone signal.

According to the SNR definition in (2.2), an orthogonal transform does not change the SNR, i.e., the SNR in the transform domain is exactly equal to the SNR in the time domain. This is because of the energy preservation property of orthogonal transforms. This implies that the SNR of the signal in the frequency domain in Fig. 1(b) is still -11dB. However, one can clearly see the signal in the frequency domain. This suggests that the SNR definition in (2.2) is not proper to judge the possibility of detecting the signal in the frequency domain in Fig. 1(b). It should not be surprising since the signal in Fig. 1(b) is not stationary and the mean power over the whole frequency domain is, of course, not proper to the signal with a single spike.

The above observation suggests that the SNR definition is transform domain dependent and should relate to the bandwidth of a signal occupied in that domain. We now introduce the following SNR definition in a domain.

Suppose the expression (2.1) is already in a transform domain, where n is the discrete variable in the transform domain. Assume the additive white noise $\eta[n]$ in (2.1) occupies the full band in the transform domain. For the signal $x[n]$ of length N, $0 \le n \le N-1$, let

$$\mathcal{B} \triangleq \{n:\ 0 \le n \le N-1 \text{ and } |x[n]|^2 \ge 0.5 \max_{0 \le n \le N-1} |x[n]|^2\}, \tag{2.3}$$

where the number 0.5 comes from the common 3dB bandwidth definition in communications. Then, the SNR is defined as

$$\text{SNR} \triangleq \frac{\sum_{n \in \mathcal{B}} |x[n]|^2}{|\mathcal{B}|\sigma^2}, \tag{2.4}$$

where $|\mathcal{B}|$ denotes the cardinality of the set $\mathcal{B}$. Notice that this definition is similar to the SNR definition in communications, where the signal is only considered in its bandwidth.

One can clearly see that the SNR in (2.4) is always greater than or equal to the SNR in (2.2) because the mean in (2.4) is only taken over the first large values in the whole domain. With the SNR definition in (2.4), the SNR in the time domain for the signal in Fig. 1(a) is -8.4dB but the SNR in the frequency domain for the signal in Fig. 1(b) is 16.3dB. Although about 2.6dB SNR is increased over the original definition in (2.2), the SNR in the frequency domain is significantly better than the old SNR, that is -11dB, in describing the signal characteristics over the noise. The time domain SNR increase is consistent for relatively stationary signals without dramatic jumpings in the time domain.

3 Signal-to-Noise Ratio in the Joint Time-Frequency Domain

In this section, we analyze the SNR in the joint time-frequency domain for the short-time Fourier transform, where the SNR defined in (2.4) is used. In order to do so, we first describe a multicomponent signal model.

3.1 Multicomponent Signal Model

Throughout the rest of this paper, we use the following multicomponent signal model,

$$y(t) = \sum_{k=1}^{K} x_k(t) + \eta(t),\ 0 \le t \le T_0, \tag{3.1}$$

where we have the following assumptions

(i) t is the continuous-time variable and limited in the finite observation interval $[0, T_0]$.

(ii) $\eta(t)$ is an additive white noise process with mean 0 and variance σ^2. It is not differentiable at any time $t \in [0, T_0]$ and independent of $x_k(t)$, $1 \le k \le K$.

(iii) For each k, $1 \le k \le K$, $x_k(t)$ is a monocomponent time-varying signal, i.e.,

$$x_k(t) = A_k(t) e^{j\phi_k(t)}, \tag{3.2}$$

where $A_k(t)$ is the slowly varying amplitude envelope of $x_k(t)$, and $\phi_k(t)$ is the phase of $x_k(t)$. The magnitude of the first order derivative $A_k'(t)$ is upper bounded by A_k, i.e., $|A_k'(t)| \le A_k$

for a positive constant A_k, and the magnitude of the second order derivative $\phi_k''(t)$ is also upper bounded by ϕ_k, i.e., $|\phi_k''(t)| \le \phi_k$ for a positive constant ϕ_k for all $t \in [0, T_0]$.

(iv) The K instantaneous frequencies $\phi_k'(t)$, $1 \le k \le K$, are distinct.

Additional details on multicomponent signals can be found in [8]. It can be easily shown that the process $y(t)$ in (3.1) has locally stationary behavior [9-10] in the following sense

$$|R_{yy}(t+u, s+u) - R_{yy}(t,s)| \le C|u|, \tag{3.3}$$

for a positive constant C, where R_{yy} denotes the autocorrelation function of $y(t)$.

As a remark, the nondifferentiability assumption (ii) of $\eta(t)$ makes sense. An example of such processes is the Wiener process, see for example [13]. This assumption implies that any sampled segment of $\eta(t)$ in any time interval is a white noise and has flat Fourier spectrum.

3.2 Short-Time Fourier Transform for Multicomponent Signals and SNR Calculations

For each monocomponent signal $x_k(t)$ in (3.1), by (i)-(iii) it can be shown that there exists $\epsilon_k > 0$ such that, for any $s \in (\epsilon_k, T_0 - \epsilon_k)$,

$$x_k(s+t) \approx A_k(s) e^{j(\phi_k(s) + \phi_k'(s)t)}, \ t \in [-\epsilon_k, \epsilon_k],$$

where the linear term $A_k'(s)t$ of t does not appear because of the "slowly varying" assumption in (iii) on the amplitude envelope $A_k(t)$. Since we have only finite many monocomponent signals $x_k(t)$ in (3.1), there exists $\epsilon = \min\{\epsilon_k, \ 1 \le k \le K\} > 0$ such that, for any $s \in (\epsilon, T_0 - \epsilon)$ and any k, $1 \le k \le K$,

$$x_k(s+t) \approx A_k(s) e^{j(\phi_k(s) + \phi_k'(s)t)}, \ t \in [-\epsilon, \epsilon], \tag{3.4}$$

where ϵ depends on the constants T_0, A_k, ϕ_k, $1 \le k \le K$.

With (3.4), at each time $s \in (\epsilon, T_0 - \epsilon)$ we apply N point discrete Fourier transform (DFT) for the signal $y(t)$ for $t \in (s - \frac{T}{2}, s + \frac{T}{2}]$ with the sampling rate N/T for $T = 2\epsilon$. For convenience, we assume N is even. The DFT is

$$P_y[m,l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} y((m+q)\frac{T}{N}) e^{-\frac{2\pi j q l}{N}}, \ 0 \le l \le N-1, \tag{3.5}$$

where m is in the range such that $(m - N/2 + 1)T/N \ge 0$ and $(m + N/2)T/N \le T_0$, i.e.,

$$\frac{N-2}{2} \le m \le (\frac{T_0}{T} - \frac{1}{2})N.$$

The above P_y can be decomposed into

$$P_y[m,l] = \sum_{k=1}^{K} P_{x_k}[m,l] + P_\eta[m,l], \ 0 \le l \le N-1, \ \frac{N-2}{2} \le m \le (\frac{T_0}{T} - \frac{1}{2})N, \tag{3.6}$$

where $P_{x_k}[m,l]$ and $P_\eta[m,l]$ are defined for $x_k(t)$ and $\eta(t)$:

$$P_{x_k}[m,l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} x_k((m+q)\frac{T}{N}) e^{-\frac{2\pi jql}{N}},\ 0 \le l \le N-1, \tag{3.7}$$

$$P_\eta[m,l] = \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} \eta((m+q)\frac{T}{N}) e^{-\frac{2\pi jql}{N}},\ 0 \le l \le N-1, \tag{3.8}$$

Since $\eta(t)$ is a white noise process, for each m the Fourier spectrum $E(|P_\eta[m,l]|^2)$ are flat over the whole frequency domain $0 \le l \le N-1$ as mentioned in Section 3.1. This implies that, the mean power of the noise spectrum $P_\eta[m,l]$ is also σ^2, which is the same as in the time domain.

We next want to study the mean power of $P_{x_k}[m,l]$ for the signal. Using (3.4),

$$\begin{aligned} P_{x_k}[m,l] &\approx \frac{1}{\sqrt{N}} \sum_{q=-N/2+1}^{N/2} A_k(m\frac{T}{N}) e^{j\left\{\phi_k(m\frac{T}{N})+\phi_k'(m\frac{T}{N})(m+q)\frac{T}{N}-\frac{2\pi ql}{N}\right\}} \\ &= \frac{1}{\sqrt{N}} A_k(m\frac{T}{N}) e^{j\left\{\phi_k(m\frac{T}{N})+\phi_k'(m\frac{T}{N})m\frac{T}{N}\right\}} \sum_{q=-N/2+1}^{N/2} e^{jq\frac{\phi_k'(m\frac{T}{N})T-2\pi l}{N}}. \end{aligned}$$

Therefore,

$$|P_{x_k}[m,l]| \approx |A_k(m\frac{T}{N})|\sqrt{N}\delta\left(l - \frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right). \tag{3.9}$$

By the assumption of distinct instantaneous frequencies $\phi_k'(m\frac{T}{N})$ for $1 \le k \le K$, the Fourier power spectrum $|P_{x_k}[m,l]|^2$ are located at K different frequencies $\phi_k'(m\frac{T}{N})T/(2\pi)$, $1 \le k \le K$. This implies

$$\begin{aligned} \left|\sum_{k=1}^{K} |P_{x_k}[m,l]|\right|^2 &\approx N\left|\sum_{k=1}^{K} A_k(m\frac{T}{N})\delta\left(l - \frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right)\right|^2 \\ &\approx N\sum_{k=1}^{K}\left|A_k(m\frac{T}{N})\right|^2 \delta\left(l - \frac{\phi_k'(m\frac{T}{N})T}{2\pi}\right). \end{aligned} \tag{3.10}$$

Therefore, for each fixed time $s = m\frac{T}{N}$, in the frequency domain,

$$\max_{0\le l\le N-1}\left|\sum_{k=1}^{K} |P_{x_k}[m,l]|\right|^2 \ge N\sum_{k=1}^{K}\left|A_k(m\frac{T}{N})\right|^2. \tag{3.11}$$

Now, let us come back to the time domain signal $y(m\frac{T}{N})$. The noise mean power is σ^2. The signal power at each time $t = m\frac{T}{N}$ is

$$\left|\sum_{k=1}^{K} x_k(m\frac{T}{N})\right|^2 \le \left(\sum_{k=1}^{K}\left|A_k(m\frac{T}{N})\right|\right)^2 \le K\sum_{k=1}^{K}\left|A_k(m\frac{T}{N})\right|^2. \tag{3.12}$$

By comparing (3.11) with (3.12), it is clear that the following relationship between the SNR_{tf} in the joint time-frequency domain of (3.6) and the SNR_t in the time domain of (3.1) at the sampling points $m\frac{T}{N}$:

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \ge 0.5\frac{N}{K}, \tag{3.13}$$

where 0.5 comes from the SNR definition in (2.3)-(3.4). Therefore, as the window size T is small enough

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \geq O\left(\frac{N}{K}\right). \tag{3.14}$$

Notice that the assumption of small enough window size T is equivalent to the assumption of fast enough sampling rate N/T. The derivation of (3.14) implies the following theorem.

Theorem 1 *For a multicomponent signal with K many monocomponents, the SNR in the joint time-frequency domain with the short-time Fourier transform with the rectangular window of size T, and the sampling rate N/T, increases over the SNR in the time domain on the order of $O(N/K)$ when the sampling rate is fast enough. Given the number K, this increase rate $O(N/K)$ is optimal.*

Proof: The first part has been proved by the above argument. The optimality can be proved by taking $A_k(t) = 1$ and $\phi_k(t) = c_k t^2$ for proper constants $c_k \neq 0$ for $1 \leq k \leq K$, and noticing that the inequalities in (3.9)-(3.12) become equalities in this case. □

4 Numerical Example

For simplicity in computations, we choose the following two-component signal model

$$y(t) = e^{j8\pi t^2} + e^{j\pi t^{2.5}} + \eta(t),\ 0 \leq t \leq 2, \tag{4.1}$$

where $\eta(t)$ is an additive white Gaussian noise with mean 0 and variance $\sigma^2 = 9$. The window size for the short-time Fourier transform is 1/8. The following constant of the SNR increase rate in terms of the number of points N of the DFT is illustrated in Fig. 2:

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \Big/ \frac{N}{K}. \tag{4.2}$$

One can see that, for this particular signal,

$$\frac{\mathrm{SNR}_{tf}}{\mathrm{SNR}_t} \to 0.55\frac{N}{K}, \quad \text{as } N \to \infty. \tag{4.3}$$

From Fig. 2, one can also see that the constants of the SNR increase rate have large variance when the sampling rate is not large enough but almost become invariant when the sampling rate becomes large.

5 Conclusion

In this paper, we have quantitatively analyzed the SNR increase rate in the joint time-frequency domain with the short-time Fourier transform over the SNR in the time domain for multicomponent signals in additive white noise. We have shown that the rate of the SNR increase is on the order of $O(N/K)$, where K is the number of monocomponents in a signal, N/T is the sampling rate and T is the window size in the short-time Fourier transform. Although we have presented quantitative analysis for the short-time Fourier transform with rectangular window functions, we believe that the result also holds for Gaussian window functions.

Acknowledgement

The author would like to thank Shie Qian and Victor Chen for having many stimulus and insightful discussions on the subject. He also wishes to thank the referees for their careful readings of this manuscript.

References

[1] V. Chen, "Reconstruction of inverse synthetic aperture radar image using adaptive time-frequency wavelet transform," (invite paper), *SPIE Proc. Wavelet Applications*, vol.2491, pp.373-386, 1995.

[2] M. Sun, S. Qian, X. Yan, S. B. Baumman, X.-G. Xia, R. E. Dahl, N. D. Ryan, and R. J. Sclabassi, Time-frequency analysis and synthesis for localizing functional activity in the brain, *Proceedings of the IEEE*, Special Issue on Time-Frequency Analysis, Sept. 1996.

[3] W. Williams, H. P. Zaveri, and J. C. Sackellares, "Time-frequency analysis of electrophysiology signals in epilepsy," *IEEE Trans. Engr. Med. Biol.*, pp.133-143, March/April, 1995.

[4] S. Kadambe, "The application of time-frequency and time-scale representations for speech analysis," Ph.D. dissertation, Dept. Elec. Eng., Univ. of Rhode Island, Kingston, RI, 1991.

[5] K. M. Wong and Q. Jin, "Estimation of the time-varying frequency of a signal: the Cramer-Rao bound and the application of Wigner distribution," *IEEE Trans. on Signal Processing*, vol.38, pp.519-536, March, 1990.

[6] P. Rao and F. J. Taylor, "Estimation of instantaneous frequency using the Wigner distribution," *Electronics Letters*, vol.26, pp.246-248, Feb., 1990.

[7] P. Flandrin, "A time-frequency formulation of optimal detection," *IEEE Trans. on Acoust. Speech, and Signal Proc.*, vol.36, pp.1377-1384, 1988.

[8] B. Boashash, "Time-frequency signal analysis," *Advances in Spectrum Analysis and Array Processing* (Ed. by S. Haykin), vol.I, Prentice-Hall, Englewood Cliffs, New Jersey, 1991.

[9] S. Mallat, G. Papanicolaou, and Z. Zhang, "Adaptive covariance estimation of locally stationary processes," preprint, 1995.

[10] D. L. Donoho, S. Mallat, and R. von Sachs, "Estimating covariance of locally stationary processes: consistency of best basis methods," *Proceedings of the IEEE-SP Int. Symp. on Time-Freq. Time-Scale Anal.*, pp.337-340, June, Paris, 1996.

[11] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, Englewood Cliffs., New Jersey, 1995.

[12] S. Qian and D. Chen, *Joint Time-Frequency Analysis*, Prentice-Hall, New Jersey, 1996.

[13] A. N. Shiryayev, *Probability*, Springer-Verlag, New York, 1979.

Figure Captions

Fig. 1 : Single tone signal.

Fig. 2 : SNR increase rate.

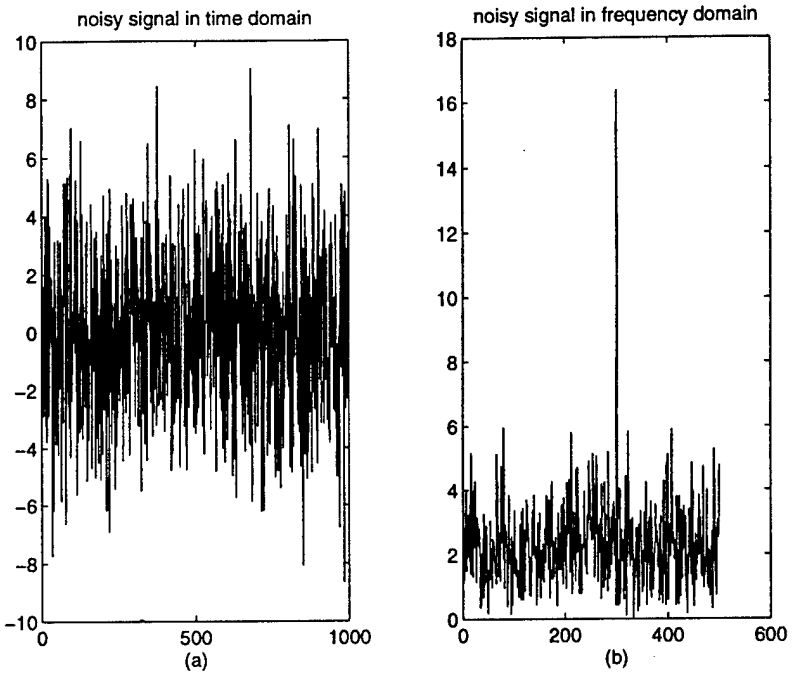

Figure 1: Single tone signal.

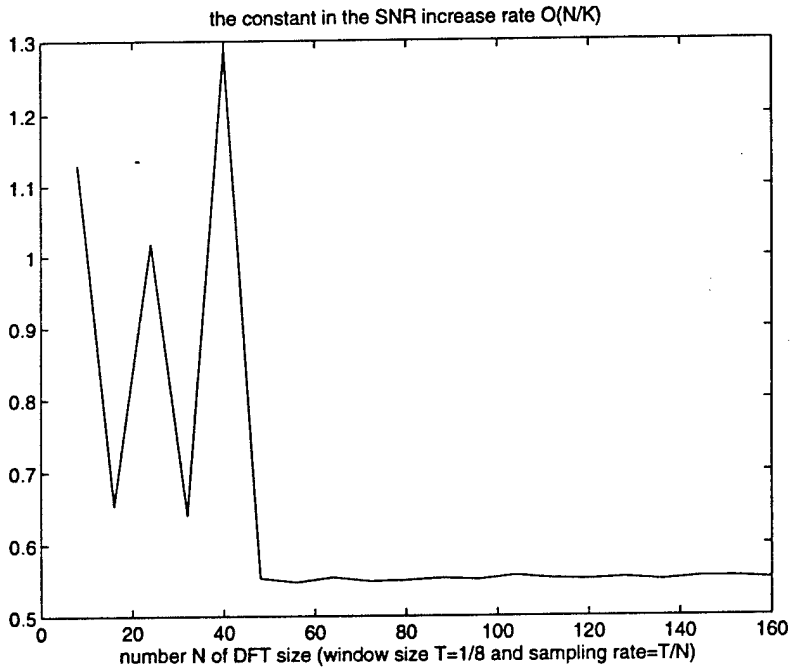

Figure 2: SNR increase rate.

Convergence of An Iterative Time-Variant Filtering Based on Discrete Gabor Transform

Xiang-Gen Xia* Shie Qian†

Abstract

An iterative time-variant filtering based on discrete Gabor transform (DGT) has been recently proposed by the authors. In this paper, we present a proof of the convergence of the iterative algorithm under a sufficient condition on the analysis and synthesis window functions of the DGT. In the meanwhile, we show that the iterative algorithm refines the least squares solution.

1 Introduction

Time-frequency (TF) transforms (or analysis) add redundancy in the joint TF domain to the signal in the time domain. They spread noise over the whole TF plane and meanwhile contain the signal information in some localized areas as shown in Fig. 1(a)-(c). Therefore, TF transforms usually significantly increase the signal-to-noise ratio in the TF domain, see for example [19] for a quantitative analysis. In other words, signals in the TF domain may be easier to be detected than in the time domain alone. With this observation, one might use the following idea for extracting the signal in the time domain analogous to traditional linear filtering: take the TF transform of a noisy signal $s(t)$; mask the TF transform in the TF plane as shown in Fig. 1(d); take the inverse TF transform of the masked TF transform shown in Fig. 1 (d) as $\tilde{s}(t)$. With traditional linear filtering, there is no question about that the Fourier spectrum of the filtered signal $\tilde{s}(t)$ has the desired frequency band. This is because the Fourier transform is a one-to-one and onto mapping for finite energy signals. Any signal in the frequency domain corresponds to a unique signal in the time domain. This is, however, no longer true in general for TF transforms. Not every signal in the joint time-frequency domain corresponds to a signal in the time domain due to the fact that TF transforms are redundant and not onto. This implies that the TF transform of the filtered $\tilde{s}(t)$

*Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716. Email: xxia@ee.udel.edu; Phone: (302)831-8038. His work was partially supported by the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, the National Science Foundation CAREER Program under Grant MIP-9703377, the 1998 Office of Naval Research (ONR) Young Investigator Program (YIP) under Grant N00014-98-1-0644, and the University of Delaware Research Foundation.

†DSP Group, National Instruments, Austin, TX 78730-5039. Email: qian@natinst.com; Phone: (512)794-5504.

may not fall in the masked domain as shown in Fig. 1 (d)-(e). With this observation, let us state the general time-frequency synthesis problem (also known as the problem of filtering, time-varying, nonstationary wideband signals). Given a user specified, localized time-frequency domain in the TF plane, find the corresponding time domain waveform. The traditional approach to this problem is the least squares solution method, which finds the signal in the time domain that minimizes the squared error between the signal's TF transform and the desired one (see, for example, [1] for ambiguity functions, [2-4] for TF transforms). For other approaches, see, for example, [5]. There are two drawbacks to the least squares solution method. The first one is that although the error between the TF transform of the solution and the desired one is minimized in the mean squared error sense, the TF transform of the solution is not guaranteed to have the desired time-frequency characteristics. This means the solution may not be the desired one as illustrated later by examples. As a result, the performance is limited, which will be seen from our numerical results later. The second drawback is the computational complexity when signals are fairly long, which is quite often the case in practice. This is because the calculation of the pseudo inverses of the matrices needed for the least squares solution method is computationally expensive when their sizes are large. Recently, an iterative time-variant filtering method based on discrete Gabor transform has been proposed by the authors, see for example [6, 7, 18].

In this paper, we present a proof of the convergence of the iterative algorithm proposed in [6, 7] under a sufficient condition on the window functions of discrete Gabor transforms. We also prove that, under these conditions, the first iteration of the iterative algorithm is exactly the least squares solution. Improvement over the least squares solution occurs with more iterations, which can be seen clearly from our numerical examples. This paper is organized as follows. In Section 2, we first briefly review discrete Gabor transforms, then restate the iterative algorithm for time-varying filtering proposed in [6, 7] and finally present a proof of the convergence. In Section 3, we present some numerical examples.

2 Convergence of the Iterative Time-Varying Filtering

In this section, we first describe the iterative time-varying filtering algorithm proposed in [6, 7] and then study its convergence.

2.1 Iterative algorithm

Let us first review the DGT studied by Wexler and Raz [8]. Let a signal $s[k]$, a synthesis window function $h[n]$ and an analysis window function $\gamma[n]$ be all periodic with the same period L. Then,

$$s[k] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} C_{m,n} h_{m,n}[k], \tag{2.1}$$

$$C_{m,n} = \sum_{k=0}^{L-1} s[k] \gamma^*_{m,n}[k], \tag{2.2}$$

$$h_{m,n}[k] = h[k - m\Delta M] W_L^{n \cdot \Delta N \cdot k}, \tag{2.3}$$

$$\gamma_{m,n}[k] = \gamma[k - m\Delta M] W_L^{n \cdot \Delta N \cdot k}, \tag{2.4}$$

and $W_L = exp(j2\pi/L)$, $j = \sqrt{-1}$, ΔM and ΔN are the time and the frequency sampling interval lengths, M and N are the number of sampling points in the time and the frequency domains, $M \cdot \Delta M = N \cdot \Delta N = L$, $MN \geq L$ (or $\Delta M \Delta N \leq L$). The coefficients $C_{m,n}$ are called the *discrete Gabor transform* (DGT) of the signal $s[k]$ and the representation (2.1) is called the *inverse discrete Gabor transform* (IDGT) of the coefficients $C_{m,n}$.

One condition on the analysis and synthesis window functions $\gamma[k]$ and $h[k]$ obtained by Wexler and Raz [8] is the following identity:

$$\sum_{k=0}^{L-1} h[k + mN] W_L^{-nMk} \gamma^*[k] = \delta[m]\delta[n], \quad 0 \leq m \leq \Delta N - 1, 0 \leq n \leq \Delta M - 1. \tag{2.5}$$

The DGT and IDGT can be also represented in the following matrix forms. Let

$$\mathbf{C} = (C_{0,0}, C_{0,1}, \cdots, C_{M-1,N-1})^T, \quad \mathbf{s} = (s[0], s[1], \cdots, s[L-1])^T.$$

The DGT can be represented by the $MN \times L$ matrix $G_{MN \times L}$ with its $(mN + n)$th row and kth column element $\gamma^*_{m,n}[k]$. The IDGT can be represented by the $L \times MN$ matrix $H_{L \times MN}$ with its kth row and $(mN + n)$th column element $h_{m,n}[k]$. Thus,

$$\mathbf{C} = G_{MN \times L} \mathbf{s} \quad \text{and} \quad \mathbf{s} = H_{L \times MN} \mathbf{C}. \tag{2.6}$$

The condition (2.5) implies that

$$H_{L \times MN} G_{MN \times L} = I_{L \times L}, \tag{2.7}$$

where $I_{L \times L}$ is the $L \times L$ identity matrix.

As mentioned in the introduction, the oversampling, which corresponds to the case when $MN > L$, of the DGT adds redundancy and is usually preferred for noise reduction applications. This can

be also seen from [19] where it is proved that the SNR in the transform domain for short-time Fourier transforms increases when the sampling rate increases. From (2.1)-(2.5), (2.6)-(2.7), one can see that an L dimensional signal $\mathbf{s}$ is transformed into an MN dimensional signal $\mathbf{C}$ and MN is greater than L due to the oversampling. Therefore, only a small set of MN dimensional signals in the TF plane have their corresponding time waveforms with length L. Let $D_{MN\times MN}$ denote the mask transform, specifically, a diagonal matrix with diagonal elements either 0 or 1. Let $\mathbf{s}$ be a signal with length L in the time domain. The first step in the time-varying filtering is to mask the TF transform of $\mathbf{s}$

$$\mathbf{C}_1 = D_{MN\times MN}G_{MN\times L}\mathbf{s},$$

where $D_{MN\times MN}$ masks a desired domain in the TF plane. Since the DGT $G_{MN\times L}$ is a redundant transformation, the IDGT of $\mathbf{C}_1$, $H_{L\times MN}\mathbf{C}_1$, may not fall in the mask. In another words, generally,

$$G_{MN\times L}H_{L\times MN}\mathbf{C}_1 \neq D_{MN\times MN}G_{MN\times L}H_{L\times MN}\mathbf{C}_1, \tag{2.8}$$

where $MN > L$, which is illustrated in Fig. 1(e). Notice that, in the critical sampling case, i.e., $MN = L$, the inequality (2.8) becomes equality. An intuitive method to reduce the difference between the right and the left hand sides of (2.8) is to mask the right hand side of (2.8) again and repeat the procedure, which leads to our iterative algorithm:

$$\mathbf{s}_0 = \mathbf{s}, \tag{2.9}$$

$$\mathbf{C}_{l+1} = D_{MN\times MN}G_{MN\times L}\mathbf{s}_l, \tag{2.10}$$

$$\mathbf{s}_{l+1} = H_{L\times MN}\mathbf{C}_{l+1}, \tag{2.11}$$

$$l = 0, 1, 2, ...$$

The above iterative algorithm is illustrated in Fig. 2.

The iterative algorithm (2.9)-(2.11) is an alternating projection procedure. It is used quite often in signal recovery applications, such as signal extrapolation and phase retrieval. The important issues for this algorithm are: When does the algorithm converge? If it converges, to what does it converge? We study these questions in the next subsection.

Before going to the convergence, let us see what the least squares solution is. Based on the definition, the least squares solution is the $L \times 1$ vector $\bar{\mathbf{x}}$ that minimizes

$$\|G_{MN\times L}\bar{\mathbf{x}} - D_{MN\times MN}G_{MN\times L}\mathbf{s}\| = \min_{\mathbf{x}}\|G_{MN\times L}\mathbf{x} - D_{MN\times MN}G_{MN\times L}\mathbf{s}\|, \tag{2.12}$$

where norm $\|\cdot\|$ is the usual Euclidean norm. Then,

$$\bar{\mathbf{x}} = (G^{\dagger}_{MN\times L}G_{MN\times L})^{-1}G^{\dagger}_{MN\times L}D_{MN\times MN}G_{MN\times L}\mathbf{s}, \tag{2.13}$$

where † stands for the complex conjugate and transpose. Clearly, when the signal length L is large, the inverse matrix computation is expensive. Although the error in (2.12) is minimized, the DGT of the least squares solution $\bar{\mathbf{x}}$ may not fall in the mask $D_{MN\times MN}$: $G_{MN\times L}\bar{\mathbf{x}} \neq D_{MN\times MN}G_{MN\times L}\bar{\mathbf{x}}$, when $MN > L$. Note that when the analysis and synthesis window functions are the same, i.e., $H_{L\times MN} = G^\dagger_{MN\times L}$, The least squares solution $\bar{\mathbf{x}}$ in (2.13) reduces to

$$\bar{\mathbf{x}} = H_{L\times MN}D_{MN\times MN}G_{MN\times L}\mathbf{s},$$

which is the first step $\mathbf{s}_1$ of the iterative algorithm (2.9)-(2.11).

2.2 Convergence of the iterative algorithm

In this subsection, we want to have a condition on the window functions $h[n]$ and $\gamma[n]$ for the convergence of (2.9)-(2.11). We show that, under this condition, the limit of the sequence $\mathbf{s}_l$ in (2.11) does have its DGT falling in the mask $D_{MN\times MN}$.

Rewrite (2.9)-(2.11) as follows

$$\mathbf{C}_{l+1} = D_{MN\times MN}G_{MN\times L}H_{L\times MN}\mathbf{C}_l = (D_{MN\times MN}G_{MN\times L}H_{L\times MN})^l D_{MN\times MN}G_{MN\times L}\mathbf{s}, \quad (2.14)$$

where $l = 1, 2, ...$ If we can prove that both matrices $D_{MN\times MN}$ and the product $G_{MN\times L}H_{L\times MN}$ are orthogonal projections, the above iterative algorithm converges by the alternating orthogonal projection theorem (see [14]). A matrix A is an *orthogonal projection* [14-15] means that (i) $A^2 = A$ and (ii) $A^\dagger = A$, and vice versa.

It is clear that the mask matrix $D_{MN\times MN}$ is an orthogonal projection. For the product matrix $G_{MN\times L}H_{L\times MN}$ we have, by (2.7),

$$(G_{MN\times L}H_{L\times MN})^2 = G_{MN\times L}H_{L\times MN}, \quad (2.15)$$

i.e., the condition (i) for an orthogonal projection is satisfied. For the Hermitian property (ii), we need the following condition on the window functions h and γ (details can be found in Appendix):

$$\sum_{l=0}^{\Delta N-1} \gamma^*[lN+k]h[lN+k+m\Delta M] = \sum_{l=0}^{\Delta N-1} h^*[lN+k]\gamma[lN+k+m\Delta M], \quad (2.16)$$

for $k = 0, 1, ..., N-1$ and $m = 0, 1, ..., M-1$.

With the above condition, the following lemmais not hard to prove.

Lemma 1 *The product matrix $G_{MN\times L}H_{L\times MN}$ is Hermitian if and only if condition (2.16) for the window functions h and γ holds .*

Proof: To prove Lemma 1, we first re-express both DGT and IDGT matrices $G_{MN\times L}$ and $H_{L\times MN}$ by taking the advantage of the forms in (2.1)-(2.4).

For $l = 0, 1, ..., M-1$, let Γ_l be the following $L \times L$ diagonal matrix

$$\Gamma_l = \text{diag}(\gamma^*[0 - l\Delta M], \gamma^*[1 - l\Delta M], \cdots, \gamma^*[L-1-l\Delta M]), \tag{2.17}$$

$W_{N\times N}$ be the N points discrete Fourier transform matrix, i.e., $W_{N\times N} = (W_N^{-mn})_{0\le m,n\le N-1}$. Let $\bar{W}_{N\times L}$ be the following $N \times L$ matrix consisting of ΔN many $W_{N\times N}$ as submatrices

$$\bar{W}_{N\times L} = (W_{N\times N}, W_{N\times N}, \cdots, W_{N\times N}). \tag{2.18}$$

Then, $G_{MN\times L}$ can be rewritten as

$$G_{MN\times L} = \begin{pmatrix} \bar{W}_{N\times L}\Gamma_0 \\ \bar{W}_{N\times L}\Gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \bar{W}_{N\times L}\Gamma_{M-1} \end{pmatrix}. \tag{2.19}$$

Similarly, the matrix $H_{L\times MN}$ can be rewritten as

$$H_{L\times MN} = (\Lambda_0\bar{W}^\dagger_{N\times L}, \Lambda_1\bar{W}^\dagger_{N\times L}, \cdots, \Lambda_{M-1}\bar{W}^\dagger_{N\times L}), \tag{2.20}$$

where Λ_l is the following $L \times L$ diagonal matrix similar to Γ_l:

$$\Lambda_l = \text{diag}(h[0 - l\Delta M], h[1 - l\Delta M], \cdots, h[L-1-l\Delta M]). \tag{2.21}$$

Therefore,

$$\begin{aligned} G_{MN\times L}H_{L\times MN} &= \left(\bar{W}_{N\times L}\Gamma_m\Lambda_n\bar{W}^\dagger_{N\times L}\right)_{0\le m,n\le M-1} \\ &= \begin{pmatrix} \bar{W}_{N\times L}\Gamma_0\Lambda_0\bar{W}^\dagger_{N\times L} & \cdots & \bar{W}_{N\times L}\Gamma_0\Lambda_{M-1}\bar{W}^\dagger_{N\times L} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \bar{W}_{N\times L}\Gamma_{M-1}\Lambda_0\bar{W}^\dagger_{N\times L} & \cdots & \bar{W}_{N\times L}\Gamma_{M-1}\Lambda_{M-1}\bar{W}^\dagger_{N\times L} \end{pmatrix}. \end{aligned} \tag{2.22}$$

For $G_{MN\times L}H_{L\times MN}$ to be Hermitian we need to have

$$\bar{W}_{N\times L}\Gamma_m\Lambda_n\bar{W}^\dagger_{N\times L} = \bar{W}_{N\times L}\Gamma^*_n\Lambda^*_m\bar{W}^\dagger_{N\times L}. \tag{2.23}$$

With the form (2.18) for $\bar{W}_{N\times L}$, the above (2.23) can be simplified as follows.

$$\bar{W}_{N\times L}\Gamma_m\Lambda_n\bar{W}^\dagger_{N\times L}$$

$$= \sum_{l=0}^{\Delta N-1} W_{N\times N}\mathrm{diag}(\gamma^*[lN+0-m\Delta M]h[lN+0-n\Delta M],\cdots,$$
$$\gamma^*[lN+N-1-m\Delta M]h[lN+N-1-n\Delta M])W^*_{N\times N}$$
$$= W_{N\times N}\sum_{l=0}^{\Delta N-1}\mathrm{diag}(\gamma^*[lN+0-m\Delta M]h[lN+0-n\Delta M],\cdots,$$
$$\gamma^*[lN+N-1-m\Delta M]h[lN+N-1-n\Delta M])W^*_{N\times N}$$

Therefore, the Hermitian property (ii) for the matrix $G_{MN\times L}H_{L\times MN}$ holds if and only if

$$\sum_{l=0}^{\Delta N-1}\gamma^*[lN+k-m\Delta M]h[lN+k-n\Delta M] = \sum_{l=0}^{\Delta N-1} h^*[lN+k-m\Delta M]\gamma[lN+k-n\Delta M], \quad (2.24)$$

for $k=0,1,...,N-1$ and $0\le m,n\le M-1$.

Since $h[n]$ and $\gamma[n]$ are periodic with period $L=M\Delta M$, the condition (2.24) is equivalent to

$$\sum_{l=0}^{\Delta N-1}\gamma^*[lN+k-m\Delta M]h[lN+k+n\Delta M] = \sum_{l=0}^{\Delta N-1} h^*[lN+k-m\Delta M]\gamma[lN+k+n\Delta M], \quad (2.25)$$

for $k=0,1,...,N-1$ and $0\le m,n\le M-1$. Notice that the difference between (2.24) and (2.25) is the difference of the signs in the front of the variable n.

The condition (2.16) can be obtained from condition (2.25). This proves Lemma 1. □

With Lemma 1 and the alternating orthogonal projection theorem, we have proved the following convergence result.

Theorem 1 *When the synthesis and the analysis window functions $h[n]$ and $\gamma[n]$ satisfy condition (2.16), the iterative algorithm (2.9)-(2.11) converges.*

There are two trivial cases where the condition (2.16) holds. The first case is the orthogonal-like case: $h[n]=\gamma[n]$ for all integer n. The second case is the critical sampling case: $\Delta M = N$. Notice that the continuous Gabor transform is never orthogonal-like unless the window functions are badly localized in the frequency domain. This, however, is not the case for the DGT. The most orthogonal-like solution was studied by Qian et. al. in [9-11]. They showed that it is possible to have the analysis window function γ very close to the synthesis window function h when h is truncated Gaussian. The error between h and γ is less than 2×10^{-6} while they are of unit energy, and therefore the error is negligible. We will see numerical results later in the next section.

We next want to see what the limit of the iterative algorithm (2.9)-(2.11) is, under the condition (2.16). Assume $\bar{\mathbf{s}}$ is the limit of the sequence $\mathbf{s}_l$ and $\bar{\mathbf{C}}$ is the limit of $\mathbf{C}_l$. Then,

$$\bar{\mathbf{C}} = D_{MN\times MN}G_{MN\times L}H_{L\times MN}\bar{\mathbf{C}} = D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}},$$

and

$$\bar{\mathbf{s}} = H_{L\times MN} D_{MN\times MN} G_{MN\times L}\bar{\mathbf{s}}.$$

We want to prove

$$G_{MN\times L}\bar{\mathbf{s}} = D_{MN\times MN} G_{MN\times L}\bar{\mathbf{s}},$$

i.e., the DGT of $\bar{\mathbf{s}}$ falls in the mask $D_{MN\times MN}$. Since $G_{MN\times L}H_{L\times MN}$ is an orthogonal projection and

$$\begin{aligned} D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}} &= G_{MN\times L}H_{L\times MN}D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}} + (I - G_{MN\times L}H_{L\times MN})D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}} \\ &= G_{MN\times L}\bar{\mathbf{s}} + (I - G_{MN\times L}H_{L\times MN})D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}}, \end{aligned} \tag{2.26}$$

we have that

$$G_{MN\times L}\bar{\mathbf{s}} \perp -(I - G_{MN\times L}H_{L\times MN})D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}},$$

where $\perp$ means orthogonal. Since $D_{MN\times MN}$ is also an orthogonal projection and

$$-(I - G_{MN\times L}H_{L\times MN})D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}} = (I - D_{MN\times MN})G_{MN\times L}\bar{\mathbf{s}},$$

we have $G_{MN\times L}\bar{\mathbf{s}} = D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}}$. This proves the following Theorem 2.

Theorem 2 *Under condition (2.16), the DGT of the limit $\bar{\mathbf{s}}$ of the iterative algorithm (2.9)-(2.11) falls in the mask $D_{MN\times MN}$, i.e.,*

$$G_{MN\times L}\bar{\mathbf{s}} = D_{MN\times MN}G_{MN\times L}\bar{\mathbf{s}}. \tag{2.27}$$

With the above result, one might ask whether it violates the known fact that an image of a TF transform of a signal in the TF plane can not be of compact support. This is because that a signal can not be time and band limited simultaneously. To answer this question, we first need to know that the above known fact is true for continuous TF transforms. Moreover, its proof is based upon the marginal properties of TF transforms. It may not be true for discrete TF transforms. In other words, discrete TF transforms may have compact support [6].

For the least squares solution $\bar{\mathbf{x}}$ in (2.13), its Gabor transform $G_{MN\times L}\bar{\mathbf{x}}$ is the orthogonal projection of $D_{MN\times MN}G_{MN\times L}\mathbf{s}$ onto the space of all signals $G_{MN\times L}\mathbf{x}$. Since $G_{MN\times L}H_{L\times MN}$ is an orthogonal projection, by (2.26), we have proved that the least squares solution

$$\bar{\mathbf{x}} = H_{L\times MN}D_{MN\times MN}G_{MN\times L}\mathbf{s} = \mathbf{s}_1.$$

This proves the following Theorem 3.

Theorem 3 *Under condition (2.16), the first iteration* $\mathbf{s}_1$ *of the iterative algorithm (2.9)-(2.11) is equal to the least squares solution in (2.13), i.e.,* $\mathbf{s}_1 = \bar{\mathbf{x}}$.

With Theorem 3, one will see in the next section that the iterative algorithm (2.9)-(2.11) improves the least squares solution when the number of iterations increases, and in the meanwhile one does not need to compute the inverse matrix in (2.13). Theorem 3 also provides another way to compute the least squares solution when condition (2.16) holds on the window functions. Note that the least squares solution in (2.13) does not depend on the synthesis window function $h[n]$. This means that all the least squares solutions are the same for all pairs of synthesis and analysis window functions as long as the analysis window functions are the same, such as the Gaussian function. Therefore, the improvement from the iterative algorithm with window functions satisfying Condition (2.16) is over the least squares solutions not only for the window functions satisfying Conditon (2.16) but also for other window functions.

3 Numerical Examples

In this section, we test two sets of window functions of the DGT. The first set of window functions are the most orthogonal-like ones obtained from [11, 18-19]. For this set of window functions, their difference, and the absolute values of the differences between the left and right hand sides of condition (2.16) are shown in Fig. 3, respectively. The second set of window functions only satisfies the Wexler-Raz condtion (2.5) and correspondingly, they are shown in Fig. 4. The test signal is $s[n] = x[n] + \eta[n]$, where $x[n] = \cos(2\pi((n+1)/115)^3)$ and $\eta[n]$ is white Gaussian noise. The mean square errors between the true signal $x(n)$ and the filtered ones are shown in Fig. 5 and Fig. 6 for the two sets of window functions, respectively. One can clearly see the performance difference.

4 Conclusion

In this paper, we presented a convergence proof of the iterative time-variant filtering algorithm proposed in [6, 7]. We proved that, under the condition, the limit of the time waveforms from the iterative algorithms has the desired TF characteristics. We also proved that, under the condition, the first iteration is equal to the least squares solution.

References

[1] C. Wilcox, "The synthesis problem for radar ambiguity functions," Tech. Summary Rep. 157, Math. Res. Center, Univ. Wisconsin, Madison, April, 1960.

[2] G. F. Boudreaux-Bartels and T. W. Parks, "Time-varying filtering and signal estimation using Wigner distribution synthesis techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol.34, pp.442-451, June, 1986.

[3] S. Farkash and S. Raz, "Time-variant filtering via the Gabor expansion," *Signal Processing V: Theories and Applications*, pp.509-512, New York: Elsevier, 1990.

[4] F. Hlawatsch and W. Krattenthaler, "Bilinear signal synthesis," *IEEE Trans. on Signal Process.*, vol.40, pp.352-363, Feb., 1992.

[5] F. Hlawatsch and W. Kozek, "Time-frequency projection filters and time-frequency signal expansions," *IEEE Trans. on Signal Process.*, vol.42, pp.3321-3334, Dec., 1994.

[6] S. Qian and D. Chen, *Joint Time-Frequency Analysis*, Prentice-Hall, New Jersey, 1996.

[7] X.-G. Xia and S. Qian, "An iterative algorithm for time-variant filtering in the discrete Gabor transform domain," *Proc. ICASSP*, pp.2121-2124, Munich, Germany, 1997.

[8] J. Wexler and S. Raz, "Discrete Gabor expansions," *Signal Processing*, vol. 21, pp.207-220, 1990.

[9] S. Qian and D. Chen, "Discrete Gabor transform," *IEEE Trans. on Signal Processing*, vol. 41, pp.2429-2438, July, 1993.

[10] S. Qian and D. Chen, "Optimal biorthogonal analysis window function for discrete Gabor transform," *IEEE Trans. on Signal Processing*, vol. 42, pp.694-697, March, 1994.

[11] S. Qian, K. Chen, and S. Li, "Optimal biorthogonal functions for finite discrete-time Gabor expansion," *Signal Processing*, vol.27, pp.177-185, 1992.

[12] A. J. E. M. Janssen, "Duality and biorthogonality for discrete-time Weyl-Heisenberg frames," RWR-518-RE-94001-ak unclassified rep. 002/94, Philips Research Laboratories, Eindhoven, the Netherlands, 1994.

[13] X.-G. Xia, "On characterization of the optimal biorthogonal window functions for Gabor transforms," *IEEE Trans. on Signal Processing*, vol.44, pp.133-136, Feb., 1996.

[14] J. von Neumann, *The Geometry of Orthogonal Spaces*, vol. II, Princeton University Press, Princeton, NJ, 1950.

[15] G. H. Golub and C. F. van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, Maryland, 1983.

[16] S. Qiu and H. G. Feichtinger, "Discrete Gabor structures and optimal representations," *IEEE Trans. on Signal Processing*, vol. 43, pp.2258-2268, Oct. 1995.

[17] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[18] X.-G. Xia, "System identification using chirp signals and time-variant filters in the joint time-frequency domain," *IEEE Trans. on Signal Processing*, vol. 45, pp.2072-2084, August 1997.

[19] X.-G. Xia, " A quantitative analysis of SNR in the short-time Fourier transform domain for multicomponent signals," *IEEE Trans. on Signal Processing*, Jan. 1998.

[20] H. G. Feichtinger and T. Strohmer, (eds), *Gabor Analysis and Algorithms*, Birkhauser, Boston, 1997.

Figure Captions

Fig. 1 TF transform illustration.

Fig. 2 Iterative time-varying filtering algorithm.

Fig. 3 The first pair of window functions.

Fig. 4 The second pair of window functions.

Fig. 5 The first set window functions: Solid line: SNR vs. iteration steps, where the least squares solution is marked by *; Dashed line: The errors between masked and unmasked DGT of the iteration solutions.

Fig. 6 The second set window functions: Solid line: SNR vs. iteration steps; Dashed line: The errors between masked and unmasked DGT of the iteration solutions. The least squares solution is marked by *.

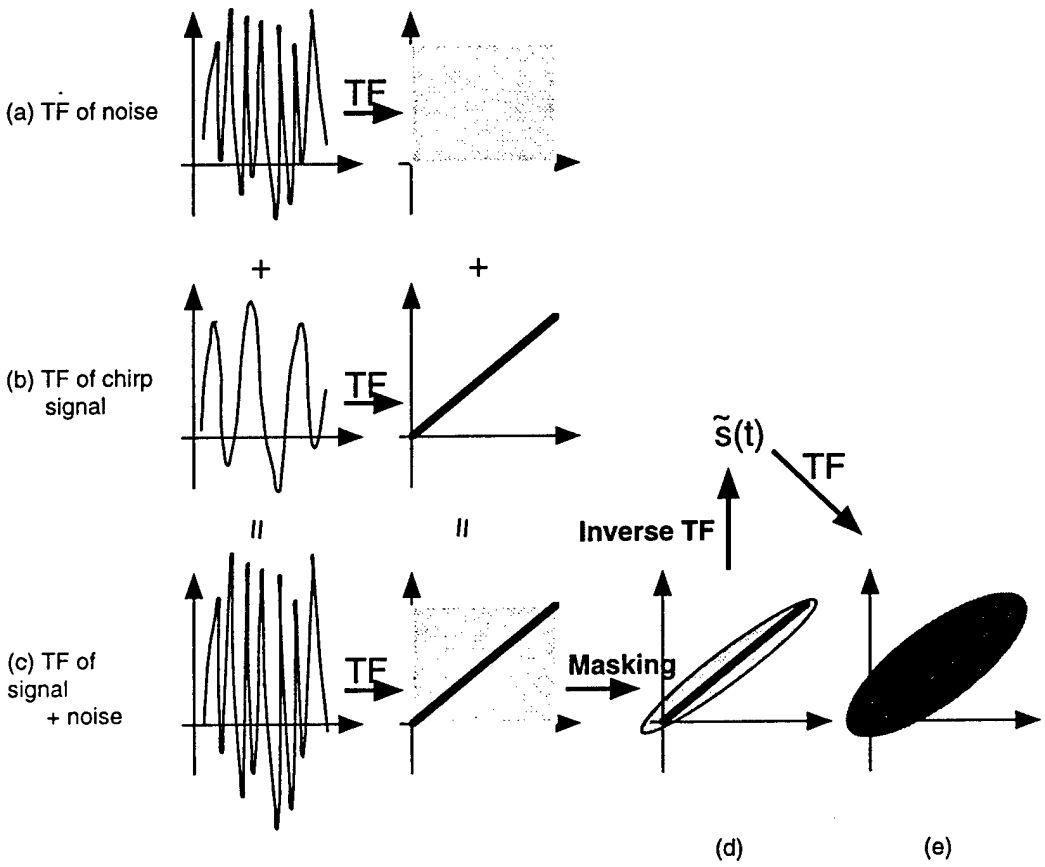

Figure 1: TF transform illustration.

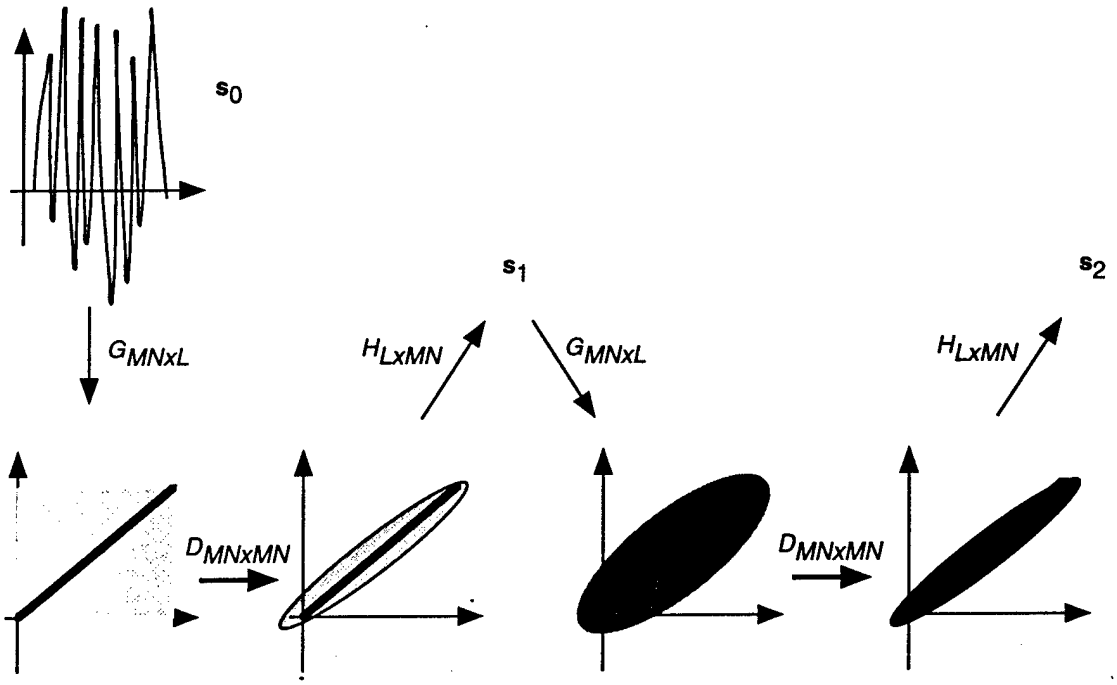

Figure 2: Iterative time-varying filtering algorithm.

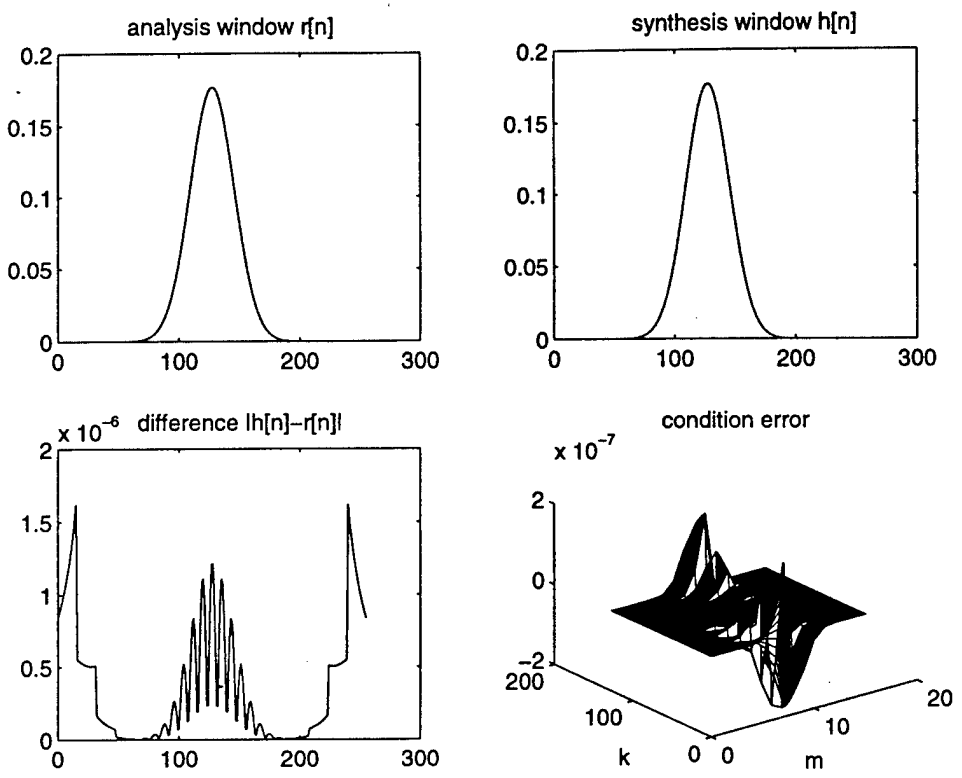

Figure 3: The first pair of window functions.

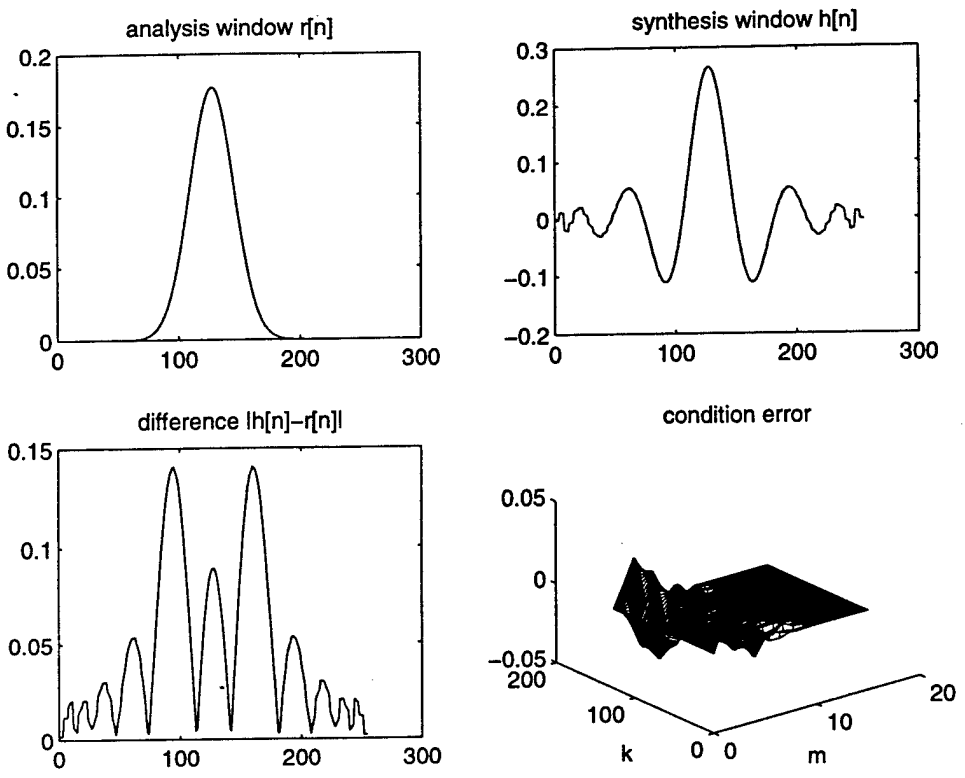

Figure 4: The second pair of window functions.

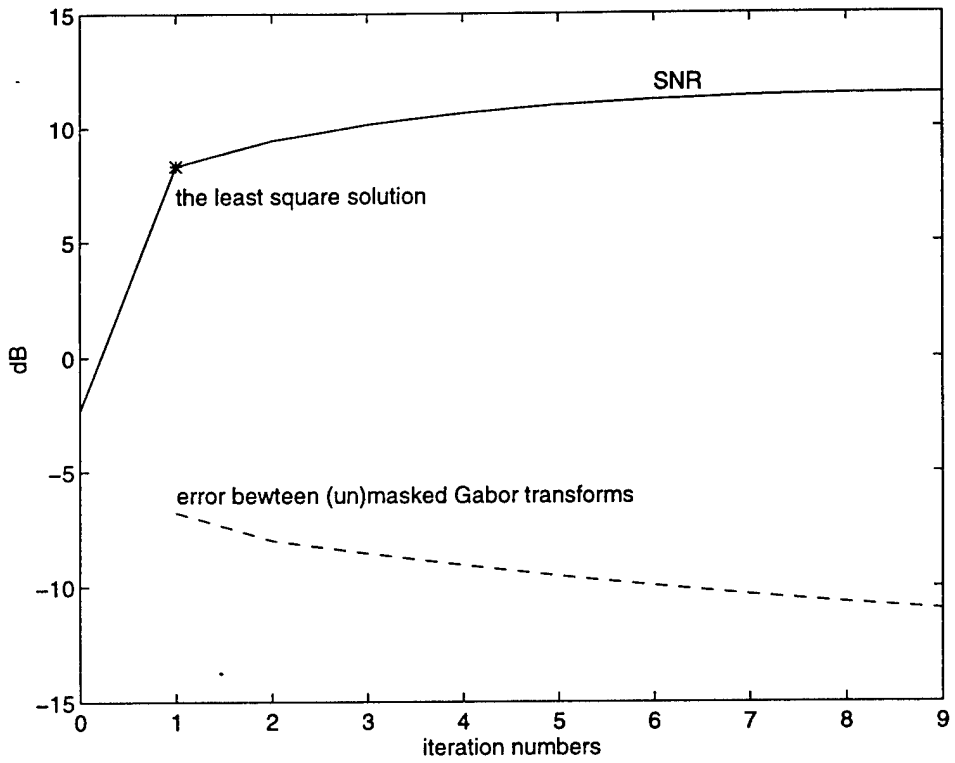

Figure 5: The first set window functions: Solid line: SNR vs. iteration steps, where the least squares solution is marked by *; Dashed line: The errors between masked and unmasked DGT of the iteration solutions.

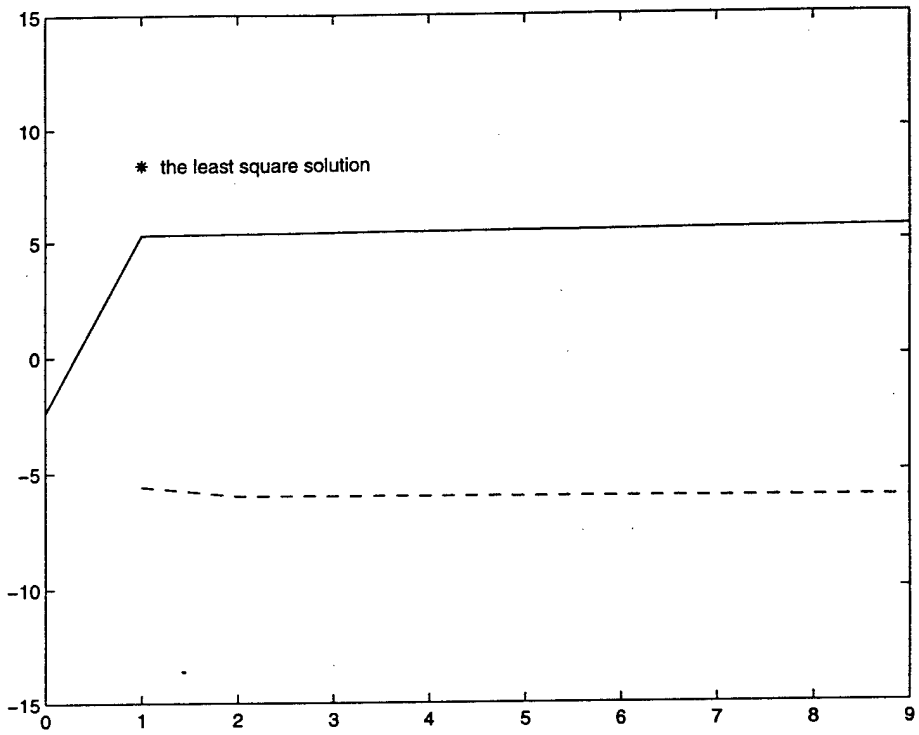

Figure 6: The second set window functions: Solid line: SNR vs. iteration steps; Dashed line: The errors between masked and unmasked DGT of the iteration solutions. The least squares solution is marked by $*$.

On Estimation of Multiple Frequencies in Undersampled Complex Valued Waveforms

Xiang-Gen Xia*

June 14, 1999

EDICS: SP2.5 (Signal Reconstruction)

Abstract

In this paper, the determination of multiple frequencies in undersampled waveforms is studied using multiple smaller size discrete Fourier transforms (DFT). Given the sizes of multiple DFT, a range for the detectable frequencies in undersampled waveforms is presented.

1 Introduction

One intuitive way to detect the single frequency f in a single frequency complex-valued waveform $x(t)$ is first to sample $x(t)$ at a sampling frequency $f_s > f$ and then to implement the N point discrete Fourier transform (DFT) with $N \geq f_s$ and a single peak in the DFT domain can be seen. The reason why the above sampling frequency f_s does not have to be at least twice of the frequency f is because the frequency of the waveform $x(t)$ is only single sided. When the frequency f is large, the sampling frequency is also large in this method. Several methods to detect a single frequency in undersampled waveforms have appeared, see for example [1-4]. The basic idea for these methods is to use multiple DFTs with smaller sizes for undersampled waveforms with different sampling rates. One of the advantages of

*Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716. Email: xxia@ee.udel.edu; Phone: (302)831-8038. His work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377, the 1998 Office of Naval Research Young Investigator Program (YIP) under Grant N00014-98-1-0644, and the Unversity of Delaware Research Foundation.

using undersampled waveforms is the hardware cost reduction in applications [5]. In some applications, such as velocity synthetic aperture radar (VSAR) [8-9], the received signals may be of undersampled nature.

In this paper, we study the estimation of multiple frequencies from undersampled complex valued waveforms by also using multiple DFTs for undersampled waveforms with different sampling rates. Given the sizes of these DFTs (or the sampling rates) and the number of multiple frequencies, we provide a range of the detectable frequencies. Note that a different approach was studied in [12] in angle estimation.

2 Multiple Frequency Estimation

Without loss of generality, we assume that the multiple frequencies in a waveform $x(t)$ are $f_1 = N_1, f_2 = N_2, ..., f_\rho = N_\rho$ and N_1, N_2, ..., N_ρ are all nonnegative integers. For the integer frequency assumption, see for example [1]. The waveform $x(t)$ is represented by

$$x(t) = \sum_{l=1}^{\rho} A_l e^{2\pi j f_l t},$$

where A_l, $1 \leq l \leq \rho$, are ρ nonzero complex-valued coefficients. Let $f_s = m$ be the sampling frequency with a positive integer m. Then the sampled waveform is

$$x_m[n] = x(\frac{n}{m}) = \sum_{l=1}^{\rho} A_l e^{2\pi j N_l n/m}, \ n \in \mathbf{Z}. \tag{1}$$

The problem of interest is to detect the multiple frequencies N_l, $1 \leq l \leq \rho$, from the multiple undersampled $x(t)$: $x_{m_r}[n]$, $n \in \mathbf{Z}$, with $m_r <\max\{N_1, ..., N_\rho\}$ for $1 \leq r \leq \gamma$. The main idea in the following study is to implement the m_r point DFT for the undersampled waveforms $x_{m_r}[n]$, $0 \leq n \leq m_r - 1$ for $1 \leq r \leq \gamma$. From these multiple DFTs, the multiple frequencies can be detected. To study the details for a general solution, let us first see the single frequency case, i.e., $\rho = 1$ in (1). In this case, let $N \triangleq N_1$. Then

$$x_{m_r}[n] = Ae^{2\pi j N n/m_r}, \ n \in \mathbf{Z}, \ A \neq 0. \tag{2}$$

Let $N = n_r m_r + k_r$, $0 \leq k_r \leq m_r - 1$, i.e., $k_r = N \bmod m_r$, then the m_r point DFT of $x_{m_r}[n]$, $0 \leq n \leq m_r - 1$, is

$$\mathrm{DFT}_{m_r}(x_{m_r}[n]) = A\delta(k - k_r),\ 0 \leq k \leq m_r - 1. \tag{3}$$

That is, the residue $k_r = N \bmod m_r$ can be detected from the m_r point DFT of $x_{m_r}[n]$ for $1 \leq r \leq \gamma$. Therefore, to detect the frequency N becomes to determine the integer N from all the residues k_r for $1 \leq r \leq \gamma$. The following lemma tells us the range of the detectable N given $m_1, m_2, ..., m_\gamma$, which is called the dynamic range in [1].

Lemma 1 *The above single frequency N can be uniquely determined if*

$$0 \leq N < lcm\{m_1, m_2, ..., m_r\}, \tag{4}$$

where lcm denotes least common multiple.

Proof: Let $m \triangleq \mathrm{lcm}\{m_1, m_2, ..., m_r\}$. For $N \geq 0$, let S_N denote the $1 \times \gamma$ integer vector

$$S_N \triangleq (k_1(N), k_2(N), \cdots, k_\gamma(N)) \text{ with } k_r(N) = N \bmod m_r. \tag{5}$$

To prove Lemma 1, it is sufficient to prove that, for any two different integers N_1 and N_2 with $0 \leq N_1 \neq N_2 < m$, the vectors $S_{N_1} \neq S_{N_2}$. Assume this is not true, i.e., there exist two integers N_1 and N_2 with $0 \leq N_1 \neq N_2 < m$ such that $S_{N_1} = S_{N_2}$. In other words, $N_1 - N_2$ is a multiple of m_r for each $1 \leq r \leq \gamma$. It implies $N_1 - N_2 = nm$ for a nonzero integer n, which is impossible when $0 \leq N_1, N_2 < m$. This contradicts with the assumption, i.e., Lemma 1 is proved. □

Lemma 1 is basically the Chinese Remainder Theorem (CRT), see for example [7]. It is clear that the single frequency N can be found from the detected residues k_r, $1 \leq r \leq \gamma$, by using the CRT. The proof suggests another method to detect N by simply looking up the table of the vectors S_N defined in (5), which can be done in priori. As a remark, the reason

for maintaining the above proof is to motivate the following general solution for multiple frequency estimation.

We now study the multiple frequency estimation problem, where ρ frequencies appear in a waveform $x(t)$ with its undersampled versions shown in (1) with $m = m_1, m_2, ..., m_\gamma$. Let

$$k_{l,r} = N_l \bmod m_r,\ 1 \le l \le \rho,\ 1 \le r \le \gamma. \tag{6}$$

Then the m_r point DFT of $x_{m_r}[n]$, $0 \le n \le m_r - 1$, is

$$\mathrm{DFT}_{m_r}(x_{m_r}[n]) = \sum_{l=1}^{\rho} A_l \delta(k - k_{l,r}),\ 0 \le k \le m_r - 1,\ 1 \le r \le \gamma. \tag{7}$$

This tells us that the residue frequencies $k_{l,r}$ can be seen as peaks in the DFT domain, i.e., they can be detected from the m_r point DFT of $x_{m_r}[n]$. Thus, the determination of the original ρ frequencies $N_1, N_2, ..., N_\rho$ becomes the determination of the nonnegative integers $N_1, N_2, ..., N_\rho$ from their residues $k_{l,r} = N_r \bmod m_r$ for $1 \le l \le \rho$ and $1 \le r \le \gamma$. The following result gives a range of N_1, N_2, ..., N_ρ for the uniqueness of the determination.

Theorem 1 *Assume that a complex valued waveform $x(t)$ contains ρ different frequencies $f_l = N_l \ge 0$ for $1 \le l \le \rho$. Let m_r, $1 \le r \le \gamma$, be γ sampling rates in the undersampled versions $x_{m_r}[n]$ of $x(t)$ in (1) with $m = m_r$, $1 \le r \le \gamma$. Let*

$$\gamma = \eta\rho + \theta,\ 0 \le \theta < \rho, \tag{8}$$

where η is a nonnegative integer. Then the ρ frequencies $f_l = N_l \ge 0$ for $1 \le l \le \rho$ can be uniquely determined by using the m_r point DFT of $x_{m_r}[n]$ for $1 \le r \le \gamma$ if

$$max\{N_1, N_2, \cdots, N_\rho\} < \max\{m, m_1, m_2, \cdots, m_\gamma\}, \tag{9}$$

where

$$m \triangleq \begin{cases} \min_{1 \le r_1 < r_2 < \cdots < r_\eta \le \gamma} lcm\{m_{r_1}, m_{r_2}, \cdots, m_{r_\eta}\}, & \text{if } \eta > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where η is defined in (8).

Proof: If $m \leq \max\{m_1, m_2, \cdots, m_\gamma\}$, Theorem 1 is straightforward by simply using the single DFT for the maximum m_r, $1 \leq r \leq \gamma$. Therefore, in what follows we assume $m > \max\{m_1, m_2, \cdots, m_\gamma\}$.

For an integer N, let $k_{l,r}(N) = N \bmod m_r$. For nonnegative integers $N_1, N_2, ..., N_\rho$, let $S_r(N_1, \cdots, N_\rho)$ be the following set

$$S_r(N_1, \cdots, N_\rho) = \{k_{1,r}(N_1), \cdots, k_{\rho,r}(N_\rho)\}. \tag{11}$$

Let $S(N_1, \cdots, N_\rho)$ be the following product set

$$S(N_1, \cdots, N_\rho) = S_1(N_1, \cdots, N_\rho) \times \cdots \times S_\gamma(N_1, \cdots, N_\rho). \tag{12}$$

To prove Theorem 1, it is sufficient to prove the following uniqueness: if there are two sets of ρ different nonnegative integer frequencies $\{N_1, N_2, ..., N_\rho\}$ and $\{M_1, M_2, ..., M_\rho\}$ such that $S(N_1, \cdots, N_\rho) = S(M_1, \cdots, M_\rho)$, $\max\{N_1, N_2, \cdots, N_\rho\} < m$, and $\max\{M_1, M_2, \cdots, M_\rho\} < m$, then the two frequency sets are equal, i.e., $\{N_1, N_2, ..., N_\rho\} = \{M_1, M_2, ..., M_\rho\}$. We first prove that $\{N_1, N_2, ..., N_\rho\} \subset \{M_1, M_2, ..., M_\rho\}$.

By the assumption $m > \max\{m_1, m_2, ..., m_\gamma\}$ in the beginning of the proof, we know that $\eta > 1$. By the condition $S(N_1, \cdots, N_\tau) = S(M_1, \cdots, M_\tau)$ we obtain that for N_1 and each m_r there exists at least one integer denoted by y_r with $1 \leq y_r \leq \rho = \tau$ such that $N_1 - M_{y_r} = 0 \bmod m_r$ for $1 \leq r \leq \gamma$. By (8), on the other hand, γ is at least η times larger than $\tau = \rho$. This means that there are at least η many $m_{r_1}, ..., m_{r_\eta}$ with $1 \leq r_1 < r_2 < \cdots < r_\eta \leq \gamma$ such that

$$M_{y_{r_1}} = M_{y_{r_2}} = \cdots = M_{y_{r_\eta}} \triangleq M_{l_0},$$

where $1 \leq l_0 \leq \tau$. Thus, $N_1 - M_{l_0} = 0 \bmod m_{r_e}$ for $e = 1, 2, ..., \eta$. By conditions $\max\{N_1, N_2, ..., N_\tau\} < m \leq \text{lcm}\{m_{r_1}, ..., m_{r_\eta}\}$ and $\max\{M_1, M_2, ..., M_\tau\} < m \leq \text{lcm}\{m_{r_1}, ..., m_{r_\eta}\}$ from the definition of m in (10), we conclude $N_1 = M_{l_0}$ similar to the proof of Lemma 1. This proves that $N_1 \in \{M_1, M_2, ..., M_\rho\}$, and therefore $\{N_1, N_2, ..., N_\rho\} \subset \{M_1, M_2, ..., M_\rho\}$ can be

similarly proved. By the same argument, we can prove $\{M_1, M_2, ..., M_\rho\} \subset \{N_1, N_2, ..., N_\rho\}$. This proves Theorem 1. □

From the proof, similar to what was mentioned after the proof of Lemma 1, the ρ frequencies N_1, N_2, ..., N_ρ can be detected by looking up the table of the product sets $S(N_1, N_2, ..., N_\rho)$ defined in (11)-(12). The uniqueness in Theorem 1 guarantees the correctness of the solution when the condition (9) is satisfied, i.e., when these frequencies are in the range defined by (9). Other determination methods similar to the CRT for single frequency estimation might exist and are definitely interesting.

3 Example

In this section, we see one simple example. Consider the case of two frequencies N_1 and N_2. We choose $m_1 = 17$, $m_2 = 19$, $m_3 = 20$, and $m_4 = 21$. In this case, $\rho = 2$, $\gamma = 4$, and therefore $\eta = 2$ in (8). Clearly, $m = m_1 m_2 = 323$. By Theorem 1, all two different frequencies N_1 and N_2 in the range $[0, 322]$ can be uniquely determined from the undersampled waveforms with sampling rates 17, 19, 20 and 21 by using 17, 19, 20 and 21 point DFTs, respectively. We can see that the sampling rates are more than 15 times less than the Nyquist sampling rate when N_1 and N_2 are close to 322.

4 Conclusion

In this paper, we studied the estimation of multiple frequencies in undersampled complex valued waveforms using multiple DFTs. Given the sizes of these multiple DFTs or the undersampling rates, we provided a range for the detectable frequencies. Our example shows that a significant sampling rate reduction over the Nyquist sampling rate can be achieved. It should be noticed that the range determined in Theorem 1 might not be the maximal one. The search of the maximal range given ρ, m_1, ..., m_γ is under our current investigation. After this paper was written, some results on the maximal range were obtained in [10]

with a sufficient condition on the multiple frequencies. The approach in this paper might be generalized to multidimensional frequency estimation by using multidimensional Chinese Remainder Theorem in [6]. In [11], the results have been recently applied in enlarging the dynamic range of the detectable parameters for polynomial phase signals using multiple lag diversities in high-order ambiguity functions.

References

[1] P. E. Pace, R. E. Leino, and D. Styer, "Use of the symmetrical number system in resolving single-frequency undersampled aliases," *IEEE Trans. on Signal Processing*, vol. 45, pp. 1153-1160, May 1997.

[2] R. B. Sanderson, J. B. Y. Tsui, and N. Freese, "Reduction of aliasing ambiguities through phase relations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, pp. 950-955, Oct. 1992.

[3] J. L. Brown, Jr., "On the uniform sampling of a sinusoidal signal," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, pp. 103-106, Jan. 1988.

[4] C. M. Rader, "Recovery of undersampled periodic waveforms," *IEEE Trans. on Acoust. Speech, Signal Processing*, vol. 25, pp. 242-249, June 1977.

[5] G. Hill, "The benefits of undersampling," *Electron. Des.*, pp. 69-79, July 1994.

[6] Y.-P. Lin, S.-M. Phoong, and P. P. Vaidyanathan, "New results on multidimensional Chinese remainder theorem," *IEEE Signal Processing Letters*, vol. 1, pp. 176-178, Nov. 1994.

[7] J. H. McClellan and C. M. Rader, *Number Theory in Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1979.

[8] B. Friedlander and B. Porat, "VSAR: A high resolution radar system for ocean imaging," *IEEE Trans. AES.*, vol. 34(3), 1998, pp.755-776.

[9] G. Wang, X.-G. Xia, and V. C. Chen, "Multi-Frequency VSAR Imaging of Moving Targets," preprint, 1998.

[10] G. C. Zhou and X.-G. Xia, "Multiple frequency detection in undersampled complex-valued waveforms with close multiple frequencies, *Electronics Letters*, vol.33, no.15, pp.1294-1295, July 1997.

[11] X.-G. Xia, Dynamic range determination of the detectable parameters for polynomial phase signals using multiple lag diversities in high-order ambiguity functions, in Proceedings of IEEE-SP Internal. Symposium on Time-Frequency and Time-Scale Analysis, Pittsburgh, PA, Oct. 6-9, 1998.

[12] M. D. Zoltowski and C. P. Mathews, "Real-time frequency and 2-D angle estimation with sub-Nyquist spatio-temporal sampling," *IEEE Trans. on Signal Processing*, vol.42, pp.2781-2794, Oct. 1994.

Proceedings of Asilomar '97, Nov. 1997

Precoding Techniques for Undersampled Multi-receiver Communication Systems*

Hui Liu
Department of Electrical Engineering
University of Virginia
Charlottesville, VA 22903-2442

Xiang-Gen Xia
Department of Electrical Engineering
University of Delaware
Newwark, DE 19716

Abstract

In this paper, we investigate the feasibility of blind signal recovery from undersampled data collected from a plurality receivers. We show that although an undersampled communication system is not completely identifiable in general, such an obstacle can be overcome by employing proper precoding with an arbitrary amount of the bandwidth expansion in the transmitter. The main contribution of this study is the formulation of a generic framework for the undersampled systems, and the derivation of conditions for a class of filters which we term ambiguity resistant precoders.

1 Introduction

Because of its practical significance, blind identification of FIR channel has received considerable attention in the past decade in communications and signal processing [1]. To date, almost all research on blind identification deals with channel outputs that are sampled at least at the baud rate. In certain applications, a communication system may be *undersampled* with rate $1/LT$ ($L > 1$), for reasons ranging from fixed hardware to variable data rates of source signals. Clearly, perfect signal recovery is not possible in these scenarios. However, when a collection of low rate observations is available, it may be feasible to restore the source signals by combining partial information from different receivers.

In this paper, we study the application of multiple receivers in blind source recovery for undersampled communication systems. To put this into perspective, consider an M-receiver undersampled system depicted in Figure 1, where L is an integer. Since the receiver part is mathematically equivalent to a multiple input and multiple output (MIMO) system, one can at most blindly recover the input towards a matrix ambiguity. The question, then, becomes: is there any affordable way to restore the blind identifiability?

Filterbank precoding has been proposed to combat ISI channels in wireless communications [3]. The same concept has been applied by Giannakis for blind channel identification. In [4], it is shown that an FIR channel can be blindly determined with minimum redundancy introduced by precoding. Motivated by these studies, we propose to use precoding techniques to solve the blind identification problem for the undersampled system. More specifically, we study a class of *ambiguity resistant* precoders which is capable of removing the ambiguity introduced by undersampling. In the remainder of this paper, we shall denote the system in Figure 1 with rate K/N precoder as $[(K,N);(L,M)]$. A regular communication system with transmission induced redundancy can be cast into the same framework.

2 A Generic Framework

Throughout our discussion, the follow assumptions are invoked for the $[(K,N);(L,M)]$ system under consideration:

A.1: The precoding filter has dimension $N \times K$, where $N > K$;

A.2: $N/K \times M/L > 1$, *i.e.*, $NM > KL$.

A.1 is clearly required in the precoding, otherwise there will be no increase in redundancy which renders blind identification impossible. The same is true A.2 since MN/KL quantifies the overall system redundancy. Under A.1 and A.2, there are still many possible combinations of the four parameters, K, N, L, and M, which make a unified analysis difficult. The following lemma simplifies the our data model by casting any system that satisfies A.1 and A.2 into a generic framework.

Lemma 1 *Any* $[(\underline{K},\underline{N});(\underline{L},\underline{M})]$ *system with* $\underline{N} > \underline{K}$ *and* $\underline{N}\ \underline{M} > \underline{K}\ \underline{L}$ *can be cast into a generic* $[(K,N);(N,M)]$ *system with* $K < N < M$.

*This work was sponsored in part by the Air Force Office of Scientific Research (AFOSR) under Grants No. F49620-97-1-0318 and No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grants MIP-9703074 and MIP-9703377.

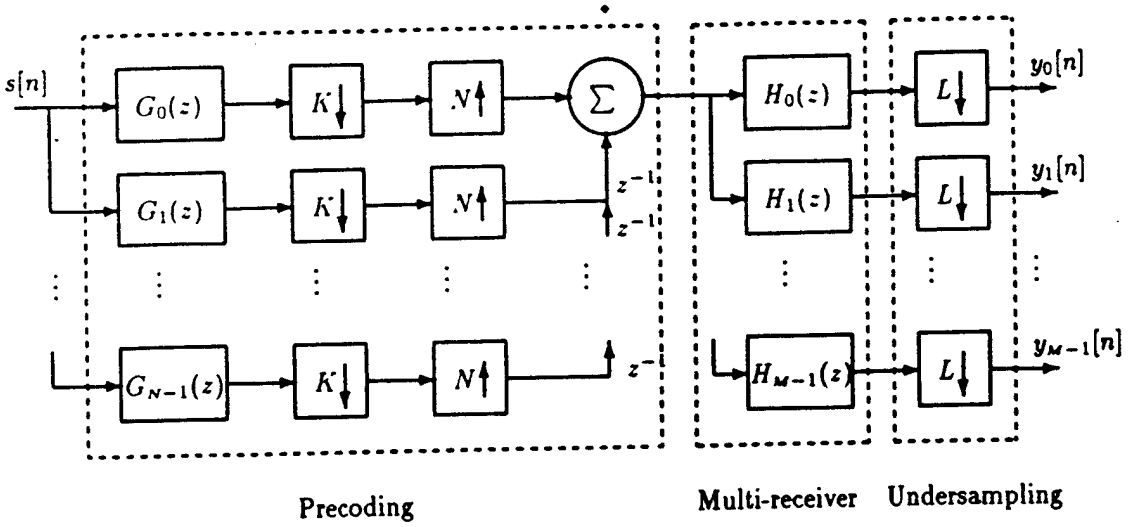

Figure 1: Precoding for undersampled an antenna array system

When $\underline{L} = \underline{M} = 1$, the system becomes a symbol-rate communication system with input redundancy. The reader is referred to [5] for proofs of the above lemma and theorems in the ensuing sections.

3 Blind Identification

The output of the generic system in Figure 2 can be expressed as

$$\mathbf{y}_{M\times 1}(z) = \mathbf{H}_{M\times L}(z)\mathbf{u}_{M\times 1}(z) = \mathbf{H}(z)\mathbf{G}_{N\times K}(z)\mathbf{s}_{K\times 1}(z),$$

where $\mathbf{H}(z)$ characterizes the *unknown* channel, whereas $\mathbf{G}(z)$ represents the *known* precoder. *The problem herein is to determine the input, $\mathbf{s}(z)$, and in many cases the channel, $\mathbf{H}(z)$, from the output, $\mathbf{y}(z)$, using only knowledge of the precoder filter, $\mathbf{G}(z)$.*

To facilitate the forthcoming discussion, let us first lay some groundwork by reviewing an important result regarding FIR MIMO systems.

Theorem 1 [6] *For an N-input M-output ($M > N$) FIR system with transfer function $\mathbf{H}(z)$, the following statements are equivalent:*

1. *$\mathbf{H}(z)$ is irreducible, i.e., $\text{rank}[\mathbf{H}(z)] = N$, $\forall z \in C$ and $\text{rank}[\mathbf{H}_o] = N$;*
2. *$\mathbf{H}(z)$ and the input vector $\mathbf{u}(z)$ can be identified up to an $N \times N$ invertible constant ambiguity matrix from the outputs using second order statistics.*

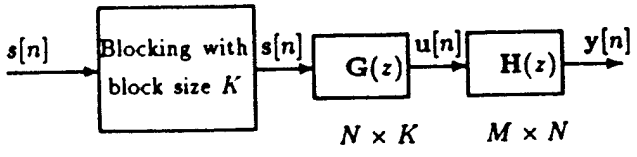

Figure 2: A generic representation

If the precoder is designed to be irreducible, the *composite* transfer function, $\mathbf{H}_c(z) \stackrel{\text{def}}{=} \mathbf{H}(z)\mathbf{G}(z)$, is clearly irreducible. Theorem 1 asserts that the system input $\mathbf{s}(z)$ can *only* be determined within a $K \times K$ matrix ambiguity directly from $\mathbf{y}(z)$. However the problem of interest here is to find $\mathbf{s}(z)$ and $\mathbf{H}_c(z)$ such that

$$\mathbf{y}(z) = \mathbf{H}_c(z)\mathbf{s}(z), \quad \text{subject to } \mathbf{H}_c(z) = \mathbf{H}(z)\mathbf{G}(z),$$

where $\mathbf{G}(z)$ is a known precoder. This motivates the following blind identifiability concept.

Definition: *The system in Figure 2 is blindly identifiable if $\tilde{\mathbf{s}}(z) = \alpha\mathbf{s}(z)$ and $\tilde{\mathbf{H}}(z) = \beta\mathbf{H}(z)$, where α and β are two scalars, are the only solution for the following system given the output $\mathbf{y}(z)$ and the precoder $\mathbf{G}(z)$: $\mathbf{y}(z) = \tilde{\mathbf{H}}(z)\mathbf{G}(z)\tilde{\mathbf{s}}(z)$.*

We tackle the blind identification problem in two steps: (i) determine what we term the *ambiguous inputs*,

$$\tilde{\mathbf{u}}(z): \quad \mathbf{T}\tilde{\mathbf{u}}(z) = \mathbf{u}(z), \tag{1}$$

where $\mathbf{T}$ is an $N \times N$ fully rank constant matrix, blindly from the system output $\mathbf{y}(z)$. It can be accomplished using many existing approaches when $\mathbf{H}(z)$ is irreducible; (ii) Once $\tilde{\mathbf{u}}(z)$ is identified, the blind identification problem reduces to whether or not $\mathbf{s}(z)$ can

be determined from $\tilde{\mathbf{u}}(z)$ in the presence of a full rank ambiguity matrix $\mathbf{T}$ (or $\mathbf{T}^{-1}$). We then show that there exists a class of *ambiguity resistant* precoders which can resolve the matrix ambiguity without additional information. Since Step (i) is well studied [6, 7], our focus in the remainder of this paper will be devoted to the precoder part.

3.1 Ambiguity Resistant (AR) Precoders

We first define the concept of ambiguity resistance.

Definition: *An $N \times K$ FIR irreducible precoding filter $\mathbf{G}(z)$ is* ambiguity resistant *if its input $\mathbf{s}(z)$ can be uniquely determined up to a scalar from its ambiguous output, $\{\tilde{\mathbf{u}}(z) : \mathbf{T}\tilde{\mathbf{u}}(z) = \mathbf{u}(z)\}$, where $\mathbf{T}$ is an unknown invertible $N \times N$ constant matrix.*

A precoder is not ambiguity resistant if there exists a non-identity matrix $\mathbf{R} \neq \alpha\mathbf{T}$ and $\tilde{\mathbf{s}}(z) \neq \beta\mathbf{s}(z)$ such that $\mathbf{R}\tilde{\mathbf{u}}(z) = \mathbf{G}(z)\tilde{\mathbf{s}}(z)$, or equivalently,

$$\mathbf{R}\mathbf{T}^{-1}\mathbf{G}(z)\mathbf{s}(z) = \mathbf{G}(z)\tilde{\mathbf{s}}(z),$$

for any given $\mathbf{s}(z)$.

Denote $\mathbf{E} = \mathbf{R}\mathbf{T}^{-1}$ and rewrite the above equation using matrix input and matrix output, the precoder is ambiguity resistant unless there exists an $N \times N$ full rank, nonidentical, constant matrix $\mathbf{E}$ and a $K \times K$ nonidentical matrix $\mathbf{X}(z)$ such that

$$\mathbf{E}\mathbf{G}(z) = \mathbf{G}(z)\mathbf{X}(z) \qquad (2)$$

$\mathbf{X}(z)$ is the *polynomial ambiguity* of the input which cannot be determined. Note that since $\mathbf{G}(z)$ is irreducible, $det(\mathbf{X}(z))$ in (2) is a nonzero constant, i.e., $\mathbf{X}(z)$ is unimodular.

The above is summarized in the following theorem.

Theorem 2 *An $N \times K$ FIR irreducible precoding filter $\mathbf{G}(z)$ is ambiguity resistant if and only if there does not exist an $N \times N$ full rank constant matrix $\mathbf{E} \neq \alpha\mathbf{I}$ for any constant α, and a $K \times K$ matrix $\mathbf{X}(z) \neq \beta\mathbf{I}$ for any constant β, such that the above identity (2) holds.*

To examine the ambiguity resistancy of a given precoder, note that it follows from Equation (2) that $\mathbf{X}(z) = \mathbf{G}^{-1}(z)\mathbf{E}\mathbf{G}(z)$. Hence,

$$\mathbf{E}\mathbf{G}(z) = \mathbf{G}(z)\mathbf{G}^{-1}(z)\mathbf{E}\mathbf{G}(z). \qquad (3)$$

By representing the above equation in the time domain, one may check the ambiguity resistancy of $\mathbf{G}(z)$ by solving a linear equation set. If $\mathbf{E} = \alpha\mathbf{I}$ for some constant α is the only nonzero solution, then $\mathbf{G}(z)$ is ambiguity resistant. Otherwise, it is necessary to check whether $\mathbf{X}(z) = \beta\mathbf{I}$ for some constant β or $\mathbf{E}\mathbf{G}(z) = \mathbf{G}(z)$, since it is possible to have $\mathbf{E}\mathbf{G}(z) = \mathbf{G}(z)$ with $\mathbf{E} \neq \alpha\mathbf{I}$.

When $K = 1$, $\mathbf{X}(z) = \alpha$ for some nonzero constant α is always true. By Theorem 2, the following corollary is straightforward.

Corollary 1 *An $N \times 1$ FIR invertible precoding filter $\mathbf{G}(z)$ is always ambiguity resistant for $N > 1$.*

Corollary 2 *Any $N \times K$ with $K > 1$ block precoder $\mathbf{G}(z)$, i.e., $\mathbf{G}(z)$ is a constant matrix, is not ambiguity resistant.*

Corollary 1 is not surprising since when $K = 1$, the $[(1, N); (N, M)]$ system reduces to a conventional oversampled system which is clearly identifiable. With this result, we only need to consider the case of $K > 1$.

Next, we want to present some necessary conditions on the ambiguity resistance.

Theorem 3 *If an $N \times K$, $K > 1$, FIR irreducible precoder $\mathbf{G}(z)$ is ambiguity resistant, then*

1. *there exist no full rank constant matrix $\mathbf{E}$ and invertible $K \times K$ polynomial matrix $\mathbf{V}(z)$ such that the first column in matrix $\mathbf{E}\mathbf{G}(z)\mathbf{V}(z)$ is $(1, 0, 0, ..., 0)^T$;*
2. *$N > K$.*
3. *the order Q of $\mathbf{G}(z)$ must satisfy the following lower bound*

$$Q \geq \frac{N^2 + K^2 - 1}{NK} - 1.$$

The above Theorem will allow us to construct a family of AR precoders in Section 3.2.

Lemma 2 *[8] When $\mathbf{G}_{N\times K}(z)$ is irreducible, its Smith-McMillan form is given by*

$$\mathbf{G} = \mathbf{W}(z)\begin{bmatrix} \mathbf{I}_{K\times K} \\ \mathbf{0} \end{bmatrix}\mathbf{V}(z),$$

where $\mathbf{W}_{N\times N}(z)$ and $\mathbf{V}_{K\times K}(z)$ are referred to as the left and right unimodular matrices, respectively, in the Smith-McMillan decomposition of $\mathbf{G}(z)$.

The left unimodular matrix, $\mathbf{W}(z)$, can be further decomposed into

$$\mathbf{W}(z) = \left[\mathbf{W}^s_{N\times K}(z)\ \mathbf{W}^c_{N\times(N-K)}(z)\right].$$

Clearly, $\mathbf{W}^s(z)$, associated with the identity part of the middle Smith form, essentially defines the column span of $\mathbf{G}(z)$. The Smith-McMillan decomposition of a tall invertible matrix can be simplied as $\mathbf{G}(z) = \mathbf{W}^s(z)\mathbf{V}(z)$, where $\mathbf{W}^s(z)$ is invertible.

Theorem 4 *A precoding filter $\mathbf{G}(z)$ is ambiguity resistant if and only if there exists no $N \times N$ constant matrix $\mathbf{E}$ such that both $\mathbf{W}(z)$ and $\mathbf{EW}(z)$, $\mathbf{W}^s(z) \neq \alpha\mathbf{EW}^s(z)$, are left unimodular matrices in the Smith-McMillan decompositions of $\mathbf{G}(z)$.*

3.2 A Family of AR Precoders

Motivated from the necessary conditions in Theorem 3, we now want to construct a family of ambiguity resistant precoders $\mathbf{G}(z)$. We have the following result.

Theorem 5 *For any positive integer $N > 1$, the following matrix $\mathbf{G}(z)$ with size $N \times (N-1)$ is ambiguity resistant:*

$$\mathbf{G}(z) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ z^{-\gamma} & 1 & 0 & \cdots & 0 & 0 \\ 0 & z^{-\gamma} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & z^{-\gamma} & 1 \\ 0 & 0 & 0 & \cdots & 0 & z^{-\gamma} \end{bmatrix}_{N\times(N-1)} \tag{4}$$

for an integer $\gamma \neq 0$.

The above theorem can be easily modified for contructing simple AR precoders of any size.

3.3 System Identifiability

With the establishment of ambiguity resistant precoders, we now give a set of sufficient conditions for blind identifiability of the system in Figure 2.

Theorem 6 *The system depicted in Figure 2 is blindly identifiable when*

1. *$\mathbf{G}(z)$ is ambiguity resistant;*
2. *$\mathbf{H}(z)$ is irreducible;*
3. *$\mathbf{G}(z)$ has order $Q \geq \lceil \frac{N(R_o+Q_h)}{K} - Q_h - R_o \rceil$, where $R_o = \lceil \frac{NQ_h}{M-N} \rceil$, and Q_h is the order of $\mathbf{H}(z)$.*

4 Algebraic Sequence Identification Algorithm

We derive in this section an algebraic algorithm which can accomplish blind identification with a finite number of observations. For this purpose we only consider noise-free data without claiming anything concerning the optimality of the algorithm.

Since the ambiguous precoder output, $\{\tilde{\mathbf{u}}[n]\}$, can be identified using one of the existing multichannel blind identification algorithms, *e.g.*, [9, 7]. we limit ourselves to the problem of removing the matrix ambiguity from $\{\tilde{\mathbf{u}}[n]\}$.

Given a finite collection of the ambiguous precoder outputs, $\{\tilde{\mathbf{u}}[n]\}_{n=0}^{R-1}$, it is not difficult to establish the following relations from (1),

$$\mathrm{diag}(\mathbf{T}\cdots\mathbf{T}) \begin{bmatrix} \tilde{\mathbf{u}}[0] \\ \vdots \\ \tilde{\mathbf{u}}[R-1] \end{bmatrix} = \mathcal{G}_R \underbrace{\begin{bmatrix} \mathbf{s}[-Q] \\ \vdots \\ \mathbf{s}[R-1] \end{bmatrix}}_{\underline{\mathbf{s}}}, \tag{5}$$

where $\mathcal{G}_R$ is a block Toeplitz matrix of $\mathbf{G}[n]$.

Upon denoting $\mathbf{t}_i$ the ith column of $\mathbf{T}$, $\underline{\mathbf{t}} = [\mathbf{t}_1^T \cdots \mathbf{t}_N^T]^T$, and $\tilde{\mathbf{U}} = [\tilde{\mathbf{u}}^T[0]\cdots\tilde{\mathbf{u}}^T[R-1]]^T \odot \mathbf{I}_{N\times N}$, we may rearrange Equation (5) with respect to its unknowns, namely, $\underline{\mathbf{s}}$ and $\underline{\mathbf{t}}$, and obtain

$$[-\mathcal{G}_R \;\; \tilde{\mathbf{U}}] \begin{bmatrix} \underline{\mathbf{s}} \\ \underline{\mathbf{t}} \end{bmatrix} = 0. \tag{6}$$

Since we have NR equations with $(R+Q)K + N^2$ unknowns, the above equation set becomes *overdetermined* as R increases, provided that $N > K$. The system can be identified using simple least squares fitting when $\mathbf{G}(z)$ is ambiguity resistant.

The above identification procedure can be summarized as follows,

1. Determine the precoder output vectors within an $N \times N$ matrix using any existing MIMO blind identification method (*e.g.*, the subspace approach in [7, 6]).
2. Form a linear equation set using the ambiguous precoder output vectors, $\{\tilde{\mathbf{u}}[n]\}_{n=0}^{R-1}$, as in (6).
3. Determine elements of the ambiguity matrix from the the least significant singular vector of (??).
4. Recover the message signals as $\mathbf{s}(z) = \mathbf{G}^{-1}(z)\mathbf{T}^{-1}\tilde{\mathbf{u}}(z)$.

5 Numerical Examples

Some numerical results are presented in this section to validate the identifiability and the efficacy of the proposed algorithms. All examples involved an $\mathbb{R}$-antenna system with the unsampling rate 3. The follow ambiguity resistant precoder described in Section 3.2 was used: $\mathbf{G}(z) = \begin{bmatrix} 1 & 0 \\ z^{-2} & 1 \\ 0 & z^{-2} \end{bmatrix}$. The system simulated is $[(2,3);(3,8)]$; and the order of the channel is 2.

The closed-form input estimation approach described in [7] was used to determine the ambiguous precoder output, $\tilde{\mathbf{u}}[n]$. Only 30 estimated vectors were applied to the proposed method. Figure 3 compares the signal constellations of the antenna outputs, the recovered precoder outputs, and the recovered signals. As shown in Figure 3 (c), existing approaches can only restore the transmitted signals, *i.e.*, the precoder outputs, within an matrix ambiguity. However, with precoding and the algorithm presented in this paper, the symbol sequence can be blindly recovery without significant increase in bandwidth.

Figure 4 shows how the mean-square error (MSE) of the symbol estimates varies with the SNR.

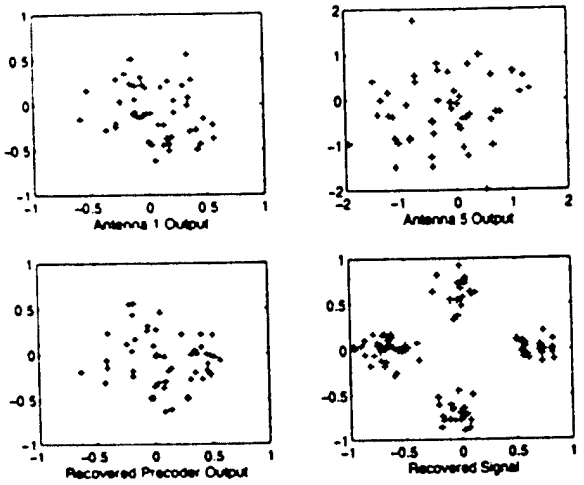

Figure 3: Signal Constellations before and after Blind Recovery

6 Concluding Remarks

In this paper, we have shown that by introducing redundancy, albeit minimum, at the input through precoding techniques, blind identification can be accomplished for undersampled systems in most scenarios. An important concept on precoders, *i.e.*, *ambiguity resistant* precoders, has been introduced and used in the blind identification. Some conditions for ambiguity resistant precoders have been given and a family of such precoders has been presented. Also presented is an algebraic algorithm which determines the unknowns of an undersampled system with a finite number of data samples.

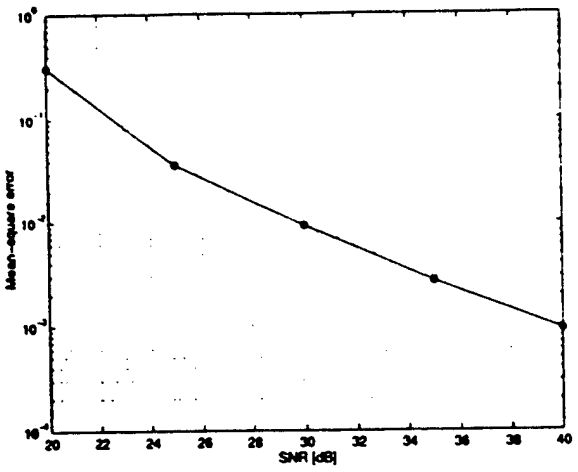

Figure 4: MSE vs. SNR

References

[1] Y. Sato, "A Method of self-recovering equalization for multilevel amplitude-modulation", *IEEE Trans. Commun.*, **23**(6):679–682, June 1975.

[2] J. G. Proakis, *Digital Communications*, McGraw-Hill Book Company, Polytechnic Institute of New York, second edition, 1989.

[3] Xiang-Gen Xia, "Intersymbol Interference Cancellation Using Nonmaximally Decimated Multirate Filterbanks with Ideal FIR Equalizers", *to appear in IEEE Trans. on Signal Processing*, 1998.

[4] G.B. Giannakis, "filterbanks for blind channel identification and equalization", *to appear in IEEE Signal Processing Letters*, 1997.

[5] Hui Liu and Xiang-Gen Xia, "Precoding techniques for undersampled multi-receiver communication systems", *submitted to IEEE Trans. on Signal Processing*, 1997.

[6] K. A. Meraim, P. Loubaton, and E. Moulines, "A subspace method for certain blind identification problems", *IEEE Trans. on Information Theory*, **IT-43**(2):499–511, 1996.

[7] H. Liu and G. Xu, "Closed-form Blind Symbol Estimation in Digital Communications", *IEEE Trans. on Signal Processing*, **SP-43**(11):2714–2723, November 1995.

[8] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[9] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters", *IEEE Trans. on Signal Processing*, **SP-43**(2):516–525, February 1995.

Proceedings of Asilomar '97, Nov. 199

Ambiguity Resistant Precoders in ISI/Multipath Cancellation: Distance and Optimality*

Xiang-Gen Xia
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716, Email: xxia@ee.udel.edu

Abstract

Ambiguity resistant (AR) precoding has been recently proposed in intersymbol interference (ISI) and multipath cancellations, where the ISI/multipath channel may have frequency-selective fading characteristics and its knowledge is not necessarily known. With the AR precoding, no diversity is necessary at the receiver. In the precoding, the AR property for a precoder plays an important rule. In this research, we introduce the concepts of precoder distance and optimal precoders, and characterize and construct all optimal systematic AR precoders, when additive channel random noise is concerned. A necessary and sufficient condition for an AR precoder to be optimal is given, which is easy to check. With the optimal precoders, numerical simulations are presented to show the improved performance over the known AR precoders in ISI cancellation applications.

1 Introduction

Intersymbol interference (ISI) and multipath fading are important problems in digital communications. Precoding is one of the techniques for the ISI/multipath cancellation. The conventional precoding techniques, such as Tomlinson-Harashima (TH) precoding and trellis precoding, and other ISI cancellation techniques, such as decision feedback equalizers, usually suffer from the spectrum-null characteristics in frequency-selective fading channels. Meanwhile, the conventional precoding methods require the knowledge of the ISI channel at the transmitter, i.e., a feedback channel is needed. Recently, a new precoding technique has been introduced in [1-6]. Unlike the conventional precoding the new precoding expands the bandwidth in a minimum amount as an expense. The advantages of the new precoding are the following: when there is no other noise but the ISI, it provides an ideal linear FIR equalizer at the receiver no matter whether or not the ISI channel has spectrum-null; it is channel independent, i.e., the feedback channel is not necessary; it is linear (no modulo operation is needed); the transmitter or receiver does not have to know the ISI channel for the equalization, i.e., blind equalization is possible.

For the blind equalization with the new precoding technique, no diversity at the receiver is needed for a single receiver system, and a reduced sampling rate over the baud rate can be achieved in an antenna array receiver system, which are not possible for the existing blind equalization techniques, see for example [7-8], without using precoding. For this purpose ambiguity resistant (AR) precoders have been introduced in [2-3] for combating the ambiguity induced by the ISI channel. Besides the AR precoder concept, some properties and families of AR precoders are presented in [2-5].

In this research, the concept of the *optimal* precoders is introduced, when additive channel random noise is concerned. The optimality is based on the following criterion: the output symbols after the precoding should be as far away from each other as possible in the mean square sense. This criterion is similar to the one in the modulation symbol design in communication systems to resist random errors. Given a precoder $\mathbf{G}(z)$, a polynomial matrix of the delay variable z^{-1}, its *distance* is introduced by using the coefficients of its coefficient matrices. It is proved that the distance is proportional to the mean distance of the ISI channel output symbols, which controls the performance in resisting additive channel random noise. Thus, an AR precoder is optimal if and only if it has the largest distance. We then characterize all optimal systematic AR precoders, where all systematic AR precoders are characterized in [4-5]. A necessary and sufficient condition for an AR precoder to be optimal is given, which is easy to check. The optimality is channel independent. Finally, Numerical examples are presented to illustrate the theory.

2 Ambiguity Resistant Precoders via ISI Cancellation

A precoded single receiver system and undersampled antenna array receiver system are shown in Fig. 1 and Fig. 2, respectively, where $\tilde{\mathbf{G}}(z)$ in Fig. 1 and $\mathbf{G}(z)$ in Fig. 2 are precoders, $H(z)$, $H_1(z)$, ..., $H_M(z)$ are the ISI channel transfer functions, and all of them are either polynomial matrices or polynomials of the delay variable z^{-1}. In what follows, boldface captial English letters denote polynomial

*This research was supported in part by an initiative grant from the Department of Electrical Engineering, University of Delaware, the Air Force Office of Scientific Research (AFOSR) under Grant No. F49620-97-1-0253, and the National Science Foundation CAREER Program under Grant MIP-9703377.

matrices.

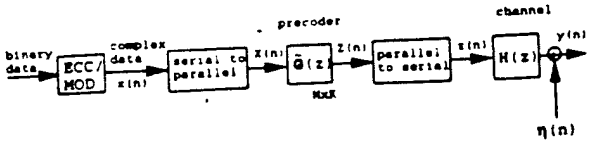

Figure 1: Single Antenna Receiver with Baud Sampling Rate.

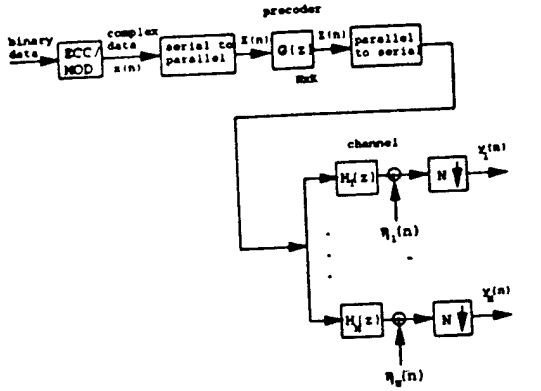

Figure 2: An Undersampled Antenna Array Receiver System.

Since the two systems in Fig. 1 and Fig. 2 can be converted to two multi-input multi-output (MIMO) systems, the existing MIMO system identification techniques, see for example [9], can be used. However, based on these results on MIMO system identification, one can at most identify an MIMO system to a constant matrix ambiguity. In order to further resist the constant matrix ambiguity induced from an MIMO system identification algorithm, ambiguity resistant precoding has been introduced in [2-3]. A precoder $\mathbf{G}(z)$ of size $N \times K$ is called *ambiguity resistant* (AR) if

(i) $\mathbf{G}(z)$ is irreducible, i.e., matrix $\mathbf{G}(z)$ has full rank for all complex values z including $z = \infty$,

(ii) the following equation for $K \times K$ polynomial matrix $\mathbf{V}(z)$ has only trivial solution $\mathbf{V}(z) = \alpha I_K$ for a nonzero constant α:

$$E\mathbf{G}(z) = \mathbf{G}(z)\mathbf{V}(z), \tag{2.1}$$

where E is an $N \times N$ nonzero constant matrix and I_K is the $K \times K$ identity matrix.

It has been shown in [2] that $\mathbf{G}(z)$ is AR implies $K < N$. In other words, the precoding has to expand each K samples into N samples. This is intuitively clear that certain redundancy is needed to resist errors. In a band-limited channel, the minimum bandwidth expansion is desired. This implies that the optimal parameter K should be $K = N - 1$ given N in an AR precoder.

Let $\tilde{\mathbf{G}}(z)$ in Fig. 1 take the following form

$$\tilde{\mathbf{G}}(z) = \begin{bmatrix} I_N \\ 0_{(M-N)\times N} \end{bmatrix} \mathbf{G}(z),$$

where $M > N$ and $\mathbf{G}(z)$ is an $N \times K$ polynomial matrix. It has been proved in [22-23] that, if the precoders in Figs. 1-2 take the above forms and $\mathbf{G}(z)$ are AR, then the input signals in the systems in Figs. 1-2 can be blindly identified from the output signals, where the ISI channels $H(z)$, $H_1(z)$, ..., $H_M(z)$ may have spectrum-null. In [2-5], families of AR precoders have been obtained. In [2-3], linear closed-form blind identification algorithms have also been obtained.

The AR precoders have been generalized in [3] to *polynomial ambiguity resistant* (PAR) precoders for resisting not only constant matrix ambiguities but also polynomial matrix ambiguities. The main advantage of using PAR precoders in the systems in Figs. 1-2 is that one can directly identify the input signals from the output signals by resolving the channel polynomial ambiguities without using any MIMO system identification algorithm. In the rest of this paper, for simplicity we however focus on AR precoders although an analogous approach applies to PAR precoders.

3 Optimal Ambiguity Resistant Precoders

Although all AR precoders found in [2-5] are good enough in theory to be used to cancel the ISI without additive noise, AR precoders may have performance difference when there is additive noise in the channel. Then the question becomes which AR precoder is "better," where "better" means better symbol error rate performance at the receiver after equalization. In this section, we study a criterion for AR precoders and also optimal AR precoders by introducing the distance concept for a precoder.

3.1 Distance and Criterion for AR Precoders

To study the above question, let us briefly recall the conventional error control coding theory. In error control coding, inputs, code coefficients and outputs are all in a finite field, such as 0 and 1, and the coding arithmetic is the finite field arithmetic. Therefore, the Hamming distance between two finite sequences of 0s and 1s is usually used. Moreover, the minimum distance between all coded sequencies can be calculated from the code itself. The minimum distance controls the performance of the error rate at the receiver for decoded sequences, when only additive random noise occurs in the channel. The differences here are, at first, the inputs, precoder coefficients and outputs are all in the complex-valued field and then the channel

has ISI besides additive random noise. Although this is the case, the "distance" of the ISI channel output values also controls the performance in resisting additive channel random noise. To the first issue the conventional Hamming distance does not apply here and the Euclidean distance for the output signal values after precoding needs to be used. Since it is hard to deal with the minimum Euclidean distance concept in the complex-valued field, the Euclidean distance here is in the mean sense when the input signal is modeled as a complex-valued random process. To the second issue, we need to investigate how the Euclidean distance of the output values of a precoder affects the Euclidean distance of the output values of the ISI channel, which determines the performance of the precoder in resisting additive random errors.

To study these issues, let us go back to the systems with ISI in Figs. 1-2. By blocking the ISI channels from serial to parallel, the systems in Figs. 1-2 can be unified into the one shown in Fig. 3, where $\mathbf{X}(z)$ is the $K \times 1$ polynomial matrix of the z-transform of the input vectors, $\mathbf{G}(z)$ is the $N \times K$ AR precoder, $\mathbf{H}(z)$ is the $M \times N$ polynomial matrix of the ISI channel, $\eta(z)$ is the $M \times 1$ polynomial matrix of the z-transform of the additive white noise vectors, and $\mathbf{Y}(z)$ is the $M \times 1$ polynomial matrix of the z-transform of the channel output vectors.

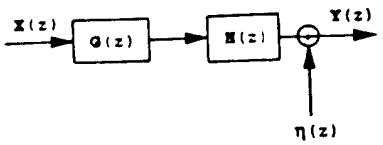

Figure 3: Unified System.

Let

$$\mathbf{G}(z) = \sum_{n=0}^{Q_G} G(n)z^{-n}, \ \mathbf{H}(z) = \sum_{n=0}^{Q_H} H(n)z^{-n},$$

$$\mathbf{X}(z) = \sum_{n} X(n)z^{-n}, \ \mathbf{Y}(z) = \sum_{n} Y(n)z^{-n}.$$

Let the z-transform of the precoder output vector sequence be

$$\mathbf{V}(z) \triangleq \mathbf{G}(z)\mathbf{X}(z) = \sum_{n} V(n)z^{-n},$$

and the z-transform of the ISI channel output vector sequence be

$$\mathbf{U}(z) \triangleq \mathbf{H}(z)\mathbf{V}(z) = \sum_{n} U(n)z^{-n}.$$

Notice that all $X(n), Y(n), V(n), V(n), \eta(n)$ are constant column vectors while $G(n), H(n)$ are constant matrices. To study the mean distance for the output values in $U(n)$, let us use matrix representations for linear transformations. By concatenating all vectors $X(n)$ together, all vectors $V(n)$ together, all vectors $U(n)$ together, all vectors $\eta(n)$ together, and all vectors $Y(n)$ together, we obtain bigger block vectors $\mathcal{X} = (x(n))$, $\mathcal{V} = (v(n))$, $\mathcal{U} = (u(n))$, $\eta = (\eta(n))$, and $\mathcal{Y} = (y(n))$, respectively. Let $\mathcal{G}$ and $\mathcal{H}$ denote the generalized Sylvester matrices, respectively:

$$\mathcal{G} = \begin{bmatrix} G(Q_G) & \cdots & G(0) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & G(Q_G) & \cdots & G(0) \end{bmatrix},$$

$$\mathcal{H} = \begin{bmatrix} H(Q_H) & \cdots & H(0) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & H(Q_H) & \cdots & H(0) \end{bmatrix}. \quad (3.1)$$

Then,

$$\mathcal{V} = \mathcal{G}\mathcal{X}, \ \mathcal{U} = \mathcal{H}\mathcal{V}, \ \mathcal{Y} = \mathcal{U} + \eta. \quad (3.2)$$

In what follows, for convenience we assume the input signal $x(n)$ is an i.i.d. random process with mean zero and variance σ_x^2. Thus, random processes $v(n)$ and $u(n)$ have mean zero. We also assume all coefficients in the ISI channel $\mathbf{H}(z)$ are i.i.d. with mean zero and variance σ_H^2 and they are independent of $x(n)$. Notice that this assumption is only used to simplify the following analysis and it does not apply to the single receiver system in Fig. 1, where the corresponding channel matrix $\mathbf{H}(z)$ has the pseudo-circulant structure [10].

The mean distances between all values of $u(n)$ and all values of $v(n)$ are

$$d_v \triangleq \left(\mathrm{E}(\sum_{m,n} |v(m) - v(n)|^2) \right)^{1/2},$$

$$d_u \triangleq \left(\mathrm{E}(\sum_{m,n} |u(m) - u(n)|^2) \right)^{1/2}, \quad (3.3)$$

respectively, where E means the expectation. By the assumptions on the coefficients of $\mathbf{H}(z)$, it is not hard to see the following relationship between the mean distance d_u of the ISI channel output values $u(n)$ and the mean distance d_v of the precoder output values (or the ISI channel input values) $v(n)$:

$$d_u = \sigma_H d_v. \quad (3.4)$$

This implies that the performance of a precoder in resisting additive channel white noise is proportional to the mean distance of the precoder output values. This result solves the second issue arised in the beginning in this section and we only need to study the mean distance d_v of all the precoder output values for the performance of resisting additive channel random errors. Based on the above analysis, we have the following definition for optimal AR precoders.

Definition 1 *An $N \times K$ ambiguity resistant precoder $\mathbf{G}(z)$ is called* optimal *if the mean distance d_v of all the precoder output values is the maximal among all $N \times K$ ambiguity resistant precoders, when the total energy is fixed.*

The squared mean distance d_v can be calculated as

$$d_v^2 = \sum_{m,n} \mathrm{E}|v(m) - v(n)|^2$$

$$= 2(LN-1)\sum_{n} \mathrm{E}(|v(n)|^2) - 2\sum_{m\neq n} \mathrm{E}(v(m)v^*(n)), \quad (3.5)$$

where L is the length of the precoder output vector sequence $V(n)$ and N is the precoder size. Let $R(m,n)$ be the correlation function of the random process $v(n)$, i.e.,

$$R(m,n) = \mathrm{E}(v(m)v^*(n)).$$

Let $\mathcal{R}$ be the correlation matrix of $v(n)$, i.e.,

$$\mathcal{R} = (R(m,n)) = \mathrm{E}\left(\mathcal{G}\mathcal{X}(\mathcal{G}\mathcal{X})^\dagger\right) = \mathcal{G}\mathrm{E}(\mathcal{X}\mathcal{X})^\dagger\mathcal{G}^\dagger = \sigma_x^2\mathcal{G}\mathcal{G}^\dagger, \quad (3.6)$$

where † means the conjugate transpose. One can see that the first term and the second term in the right hand side of (3.5) for the distance d_v are the sum of all the diagonal elements, i.e., the trace, of the matrix $\mathcal{G}\mathcal{G}^\dagger$ multiplied by $2\sigma_x^2$, and the sum of all the off diagonal elements of the matrix $\mathcal{G}\mathcal{G}^\dagger$ multiplied by $2\sigma_x^2$, respectively. In formula, the squared mean distance d_v can be calculated as

$$d_v^2 = 2\sigma_x^2\left((LN-1)\mathrm{trace}(\mathcal{G}\mathcal{G}^\dagger) - \sum_{m\neq n}(\mathcal{G}\mathcal{G}^\dagger)_{mn}\right)$$

$$= 2\sigma_x^2\left(LN\mathrm{trace}(\mathcal{G}\mathcal{G}^\dagger) - \sum_{m,n}(\mathcal{G}\mathcal{G}^\dagger)_{mn}\right), \quad (3.7)$$

where $(\mathcal{G}\mathcal{G}^\dagger)_{mn}$ denotes the element at the mth row and the nth column of $\mathcal{G}\mathcal{G}^\dagger$.

We next want to simplify d_v in (3.7) by using all the coefficients in the precoder $\mathbf{G}(z)$. For a precoder $\mathbf{G}(z)$, define

$$D_G \triangleq \text{ sum of all coefficients of all coefficient matrices of } \mathbf{G}(z)\mathbf{G}^\dagger(1/z), \quad (3.8)$$

$$E_G \triangleq \text{ sum of all magnitude squared coefficients of all coefficient matrices of } \mathbf{G}(z), \quad (3.9)$$

where $\mathbf{G}^\dagger$ means the conjugate transpose of all coefficient matrices of $\mathbf{G}(z)$. Let L be the length of the precoder output vector sequence $V(n)$. Then, by (3.1), it is not hard to see that

$$\mathrm{trace}(\mathcal{G}\mathcal{G}^\dagger) = LE_G, \text{ and } \sum_{m,n}(\mathcal{G}\mathcal{G}^\dagger)_{mn} = LD_G. \quad (3.10)$$

Therefore,

$$d_v^2 = 2\sigma_x^2 L(LNE_G - D_G). \quad (3.11)$$

Since E_G is fixed as the total energy of all the coefficients of the coefficient matrices in $\mathbf{G}(z)$, and σ_x^2, L, and N are also fixed, based on formula (3.11) for the mean distance d_v, we have the following criterion for judging the performance of an AR precoder.

Definition 2 *$N \times K$ ambiguity resistant precoder $\mathbf{G}(z)$ is said better than $N \times K$ ambiguity resistant precoder $\mathbf{F}(z)$ if $D_G < D_F$ when $E_G = E_F$, where D_G, D_F, E_G, and E_F are defined by (3.8)-(3.9) for precoders $\mathbf{G}(z)$ and $\mathbf{F}(z)$, respectively.*

Based on formula (3.11) on the mean distance d_v of the precoder output values, we define the distance for a precoder as follows.

Definition 3 *For an $N \times K$ precoder $\mathbf{G}(z)$, its distance is defined by*

$$d(G) \triangleq N - \frac{D_G}{E_G},$$

where D_G and E_G are defined in (3.8)-(3.9).

With the above two definitions the following lemma is straightforward.

Lemma 1 *AR precoder $\mathbf{G}(z)$ is better than AR precoder $\mathbf{F}(z)$ if and only if the distance of $\mathbf{G}(z)$ is greater than the distance of $\mathbf{F}(z)$, i.e., $d(G) > d(F)$.*

Since the precoder output vector length L, the precoder size N, and the input signal variance σ_x^2 are fixed, the following theorem is straightforward from (3.11).

Theorem 1 *An $N \times K$ ambiguity resistant precoder $\mathbf{G}(z)$ is optimal in all $N \times K$ ambiguity resistant precoders if and only if the total sum D_G of all the coefficients of all the coefficient matrices of the product matrix $\mathbf{G}(z)\mathbf{G}^\dagger(1/z)$ is minimal among all possible $N \times K$ ambiguity resistant precoders $\mathbf{F}(z)$ when the total sum E_G of all the magnitude squared coefficients of all coefficient matrices of $\mathbf{F}(z)$ is fixed.*

Notice that

$$\sigma_x^2 LD_G = \sigma_x^2 \sum_{m,n}(\mathcal{G}\mathcal{G}^\dagger)_{mn} = \sum_{m,n}\mathrm{E}(v(m)v^*(n))$$

$$= \mathrm{E}\left|\sum_n v(n)\right|^2 \geq 0. \quad (3.12)$$

Using (3.11), the following upper bound for the mean distance d_v is proved.

Theorem 2 *The mean distance d_v of the precoder output values for an $N \times K$ precoder $\mathbf{G}(z)$ is upper bounded by*

$$d_v \leq \sigma_x L\sqrt{2N}\sqrt{E_G}, \quad (3.13)$$

where σ_x^2 is the input signal variance, L is the length of the precoder output vector sequence, and E_G is defined by (3.9), i.e., the total energy of all coefficients in $\mathbf{G}(z)$. The upper bound for the distance of an $N \times K$ precoder $\mathbf{G}(z)$ is $d(G) \leq N$.

3.2 Optimal Systematic AR Precoders

In this subsection, we determine all optimal systematic AR precoders with the form:

$$\mathbf{F}(z) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \\ F_1(z) & F_2(z) & \cdots & F_{N-1}(z) \end{bmatrix}_{N\times(N-1)}, \tag{3.14}$$

by using the above criterion. We have the following result.

Theorem 3 *An $N \times (N-1)$ systematic ambiguity resistant precoder $\mathbf{F}(z)$ in (3.14) with*

$$F_k(z) = \sum_{l=0}^{n_k} a_{kl} z^{-l},\ a_{kn_k} \neq 0,\ 1 \leq k \leq N-1, \tag{3.15}$$

for $n_1 > n_2 > \cdots > n_{N-1} \geq 1$, is optimal if and only if

$$\sum_{l=0}^{n_k} a_{kl} = -1, \quad \text{for } k = 1, 2, ..., N-1. \tag{3.16}$$

Moreover, for the above optimal precoder, the mean distance d_v of the precoder output values and the precoder distance $d(F)$ are

$$d_v = \sigma_x L\sqrt{2N}\sqrt{E_F}, \quad \text{and} \quad d(F) = N, \tag{3.17}$$

where σ_x^2 is the variance of the input signal, L is the length of the precoder output vector sequence and

$$E_F = N - 1 + \sum_{k=1}^{N-1}\sum_{l=0}^{n_k} |a_{kl}|^2. \tag{3.18}$$

This theorem also implies that there exist AR precoders that reach the upper bound (3.13), i.e., $D_G = 0$.

4 Simulation Results and Conclusion

Some simulation results with 5 different AR precoders with different distances are shown in Fig. 4.

In this paper, we introduced the concepts of precoder *distance* and *optimal* AR precoders in justifying an AR precoder. Given an $N \times K$ precoder $\mathbf{G}(z)$, its distance is defined by $d(G) = N - D_G/E_G$, where D_G is the total sum of all coefficients of all coefficient matrices of the matrix $\mathbf{G}(z)\mathbf{G}^t(1/z)$ and E_G is the total sum of all magnitude squared coefficients of all coefficient matrices of the matrix $\mathbf{G}(z)$. With this distance definition, an $N \times K$ AR precoder is optimal if and only if its distance is N. Furthermore, we characterized all $N \times (N-1)$ optimal systematic AR precoders. With this characterization, one is able to construct all possible optimal $N \times (N-1)$ systematic AR precoders. Finally, numerical simulations were presented to illustrate the theory and the concepts. Our numerical examples showed that an optimal AR precoder has good performance in resisting both of the channel ISI and additive random noise.

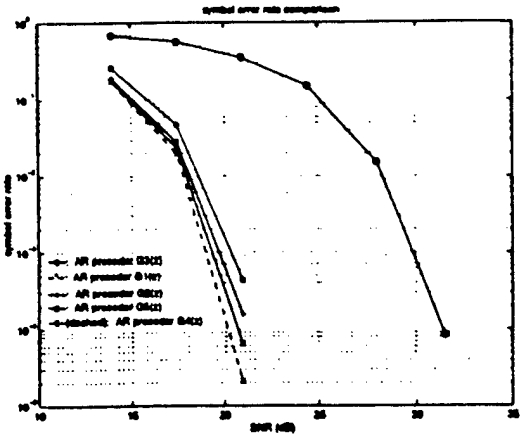

Figure 4: Symbol Error Rate Comparison: Solid line with $*$ is for $\mathbf{G}_1(z)$; Solid line with $+$ is for $\mathbf{G}_2(z)$; Solid line with o is for $\mathbf{G}_3(z)$; Dashed line with x is for $\mathbf{G}_4(z)$; Solid line with x is for $\mathbf{G}_5(z)$. $d(\mathbf{G}_1) = d(\mathbf{G}_2) = 2$, $d(\mathbf{G}_4) = d(\mathbf{G}_5) = 4$, $d(\mathbf{G}_3) = 1.0858$.

References

[1] X.-G. Xia, "New precoding for intersymbol interference cancellation using nonmaximally decimated multirate filterbanks with ideal FIR equalizers," *IEEE Trans. on Signal Processing*, vol.45, pp.2431-2441, Oct. 1997..

[2] H. Liu and X.-G. Xia, "Precoding techniques for undersampled multi-receiver communication systems," preprint, 1997.

[3] X.-G. Xia and H. Liu, "Polynomial ambiguity resistant precoders: theory and applications in ISI/multipath cancellation," preprint, 1997.

[4] X.-G. Xia and G. Zhou, "On optimal ambiguity resistant precoders in ISI/multipath cancellation," preprint, 1997.

[5] G. Zhou and X.-G. Xia, "Ambiguity resistant polynomial matrices," preprint, 1997.

[6] G.B. Giannakis, "Filterbanks for blind channel identification and equalization," *IEEE Signal Processing Letters*, June 1997.

[7] L. Tong, G. Xu, and T. Kailath, "A new approach to blind identification and equalization of multipath channel," *Proc. of the 25th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 1991.

[8] H. Liu and G. Xu, "Closed-form blind symbol estimation in digital communications," *IEEE Trans. on Signal Processing*, SP-43(11):2714-2723, November 1995.

[9] Y. Li and K. J. Ray Liu, "On blind MIMO channel identification using second-order statistics," *Proc. of 30th Conf. on Info. Scie. and Systems*, Princeton Univ., Princeton, NJ, March 1996.

[10] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice Hall, 1993.